Educational Research and Appraisal

EDUCATIONAL RESEARCH AND APPRAISAL

LIPPINCOTT SERIES IN EDUCATION

Under the editorship of Robert A. Davis

Educational Research and Appraisal

ARVIL S. BARR
PROFESSOR OF EDUCATION
UNIVERSITY OF WISCONSIN

ROBERT A. DAVIS
PROFESSOR OF EDUCATIONAL PSYCHOLOGY
GEORGE PEABODY COLLEGE FOR TEACHERS

PALMER O. JOHNSON
PROFESSOR OF EDUCATION
UNIVERSITY OF MINNESOTA

J. B. Lippincott Company
CHICAGO PHILADELPHIA NEW YORK

to some larger population. Here, inferential research, based on appropriate sampling techniques is important. Inferential research is particularly useful in a study of problems that may be expected to make some significant contribution to theory.

The book has been designed for field workers. Its principal concern is with field research and appraisal as distinguished from artificially designed laboratory studies. It stresses research that may be conducted in school settings as a basis for action. Effort has been made to present situations illustrative of the types of problem confronted by educators in their day to day activities. Teaching experience and undergraduate mastery of certain elementary concepts in measurement, statistics, and social foundations have been assumed. The early chapters are in general nonmathematical, but not necessarily less technical than those that follow.

Students who have limited statistical training should be able to understand the logic of the statistical chapters even though they may not be skilled in calculation. Those with adequate statistical training should experience little difficulty in working through the models which are in the main self-contained. Both instructors and students will find it profitable to analyze studies that illustrate the application of statistical devices in the solution of educational problems. Selected references have been supplied at the end of the book for further analysis and study.

The book was planned by Mr. Barr, who contributed Chapters V, VII, and XI. He also assisted in the writing of Chapter I and the Appendix, and in other ways served as supervisor of the project. Mr. Davis contributed Chapters II, III, and IV; a part of Chapter I; and the Appendix. Mr. Johnson prepared Chapters VI, VIII, IX, and X; and served as critic of all chapters involving the use of statistics. It was believed that a volume prepared by several persons, each having specialized interests, would ensure a sounder and more comprehensive treatment than the contribution of any one person working independently.

The authors gratefully acknowledge the assistance of a number of persons who contributed to the work of the book. Professor Max D. Engelhart of the Chicago Public Schools, Chicago, studied the typescript in its entirety and made many valuable suggestions for its improvement. Professor Julian Stanley of George Peabody College, rendered expert assistance on a number of sections in Chapter IV. Mr. Harley Ericksen, Research Assistant at the University of Wisconsin and Miss Hazel Eddins, Graduate Assistant at Peabody College, studied the book for meaningfulness to graduate students. Acknowledgment is made at appropriate places to publishers and editors who have granted permission to use their materials.

A. S. B.
R. A. D.
P. O. J.

CONTENTS

EDUCATIONAL RESEARCH AND APPRAISAL

CHAPTER I

Introduction

Growth of research in education has been one of the outstanding characteristics of cultural progress during the present century. This growth is evidenced by the continued appearance of new courses dealing with educational problems, the increased number of theses accepted each year in graduate schools, and new grants for research. In addition to these numerous evidences of interest in educational research, there are periodicals, university presses, commercial houses, and a large number of public school research bureaus that publish results of educational investigations. An analysis of studies reported in these various forms bears testimony relative to the rapid development and spread of research activities in this country.

DEVELOPMENT OF RESEARCH

Educational research in the United States has passed through several well-defined stages in its manner of solving problems. During the latter part of the 19th century educators sought answers for many of their problems through an exchange of individual experiences. Reports of many of these efforts to solve educational problems are recorded in the earlier journals. Papers and addresses of educational leaders given at educational meetings at that time consisted mainly in reporting what they were doing in their own school systems and what they had personally found effective. These papers and addresses covered a multitude of subjects, such as teaching English, managing schools, getting along with parents, and securing financial support.

This early effort at solving problems might have been labeled the personal experience method, as it provided a basis for sharing

experiences, drawing conclusions, and resolving difficulties through discussion. Although the method produced few lasting results, it had the effect of stimulating educational thinking and of preparing the way toward the scientific solution of educational problems.

The personal experience method was followed in later years by what may be termed the *deliberative* approach to solving educational problems. The deliberative approach, like the personal experience method, consisted in a discussion of problems but tended more toward committee action. Unlike the personal experience method, which at best provided only a loosely organized exchange of experience, the deliberative method provided a means of defining significant educational problems and of reaching conclusions on the basis of group thinking. This approach provided a systematic means of obtaining a consensus upon problems requiring immediate solution. The deliberative approach has continued into the present.

Era of objective measurement. Between 1910–1920, educationalists witnessed the beginning of a measurement movement destined to revolutionize appraisal and research techniques the world over. Prior to 1910 there were perhaps less than a half-dozen widely known objective tests, and these were limited to measures of general intelligence and achievement. During the decade 1910–1920, and particularly the years immediately following World War I, there were published a wide variety of instruments that became available for research and appraisal purposes. The movement has continued into the present with more and more aspects of the educational program being subjected to objective measurement and more and more people becoming specialized in its methods and techniques. *The Third Mental Measurement Yearbook* [1] published in 1949 lists 663 tests and 549 books on measurement and related subjects.

Methods of research. The second quarter of the present century can be characterized as a period during which extraordinary stress was placed upon the processes of collecting, analyzing, and quantifying educational data. *The Journal of Educational Research* and the American Educational Research Association were established in 1920. The very large amount of activity in this field made it possible for the latter organization to establish in 1930 the *Review of Educational Research,* devoted to the digest of reports of research from many sources. Many writers turned their attention to

[1] Oscar K. Buros, *The Third Mental Measurement Yearbook* (Rutgers University Press, New Brunswick, N. J., 1949).

the methods and techniques of research with the result that numerous papers, monographs, and books were published on this subject.[1]

The emphasis on practical problems. Although methods of attacking problems have changed from time to time as new instruments and techniques have become available, the principal concern of educationalists throughout this period has been the discovery of solutions to frequently observed field problems. In 1911, for example, Kohl [2] presented a paper at the Wellesley meeting of the New England Association of College Teachers of Education, in which he suggested topics that might be studied by scientific research methods. His list of problems included the following:

(1) Which is the better, one or two sessions a day?
(2) What should be the length of sessions for the different grades?
(3) Are shorter sessions for six days of the week better than longer ones for five days?
(4) Are a number of short vacations better than one or two longer ones?
(5) What seasons of the year are most conducive to good school work?
(6) What are the best days of the week for good school work?
(7) What should be the length of the school year?
(8) What hours of the day are best for hard study?
(9) How does the work of evening classes compare with that of day classes?
(10) How many studies can a pupil pursue in the different grades to his greatest profit?
(11) What should be the length of the recitation in the different grades?
(12) What is the best length of intermission for the different grades? Should it lengthen as day advances? How should the pupil use this intermission?
(13) Should studies of relatively high correlation on the content side be grouped in the same term?
(14) What is the fatigue coefficient of the different studies? Is mathematics a more fatiguing study than Latin, for exam-

[1] See Carter V. Good, A. S. Barr, and Douglas E. Scates, *Methodology of Educational Research* (D. Appleton Century Co., New York, N. Y., 1936); and Walter S. Monroe, *Encyclopedia of Educational Research* (The Macmillan Co., New York, N. Y., 1950) for lists of original publications.

[2] Clayton C. Kohl, "Needed research on the programs of studies," *J. Educ. Psych.*, 1912, *13:* 160–162.

ple? Should two or more studies apparently involving great eye strain come together?

(15) What should be the length of the lunch period?

(16) At what time may home study begin? How much and what kind of home study may be demanded?

(17) Should there be formal examination and review periods?

Many such statements of the problems of education have been made since that date. Buckingham,[1] the first editor of the *Journal of Educational Research,* in the first issue of the publication, stated that ". . . it will emphasize applications rather than abstractions, and practice rather than theory." The present editor, A. S. Barr, has pursued the same policy with frequent papers and editorials [2] on the importance of "field" or "action" research. Buckingham's *Research for Teachers* [3] provides one of the early systematic attempts to make research techniques available to the classroom teacher. Waples and Tyler [4] provided a manual for use of teachers in the study of classroom procedures.

Barr [5] in his Introduction to the *Scientific Study of Supervision* provided a manual of somewhat the same character for supervisors. A relatively recent listing of problems for further research in many aspects of education will be found in the January 1945 issue of the *Journal of Educational Research.*[6] Examination of the many issues of the *Review of Educational Research* will provide abundant evidence of the continued interest of American educators in the practical problems of field workers.

Nature of appraisal and research. During the early development of method of research and appraisal stress was placed upon the processes of collecting data by use of objective instruments; not much attention was given to controls and other essentials of a scientific methodology.

During the second and third decades of this century, there appeared numerous papers, monographs, and books discussing the

[1] B. R. Buckingham, "Announcement," *Journal of Educational Research,* 1920, 1, 1.

[2] A. S. Barr, "Research for Teachers," *Journal of Educational Research,* 1930, 20:42–43.

[3] B. R. Buckingham, *Research for Teachers* (New York: Silver, Burdett and Co., 1926).

[4] Douglas Waples and Ralph W. Tyler, *Research Methods and Teachers' Problems* (New York: Macmillan and Co., 1930).

[5] A. S. Barr, *An Introduction to the Scientific Study of Classroom Supervision* (New York: D. Appleton-Century Co., 1931).

[6] This is an anniversary issue with papers by Charters, Ashbaugh, Woody, Monroe, Brueckner, Douglass, Scates, Loomis, Symonds, Good, Barr, Buckingham, and others.

nature, logic, and methods of research. Considerable emphasis was given to the classification of research methods [1] and to descriptions of how each method might be effectively employed.[2] Progress was made in resolving some of the conflicts between the so-called historical, scientific, and philosophical approaches to research; but the close integration of the several approaches to appraisal and research were to be realized only at a much later period. In general, extraordinary progress has been made in creating a scientific attitude in the field of professional education.

Data gathering versus appraisal. Educational data are collected by use of many devices: e.g., tests, questionnaires, interviews, descriptions of behavior, and counting techniques. When these are used with due consideration for the conditions under which they can be properly administered, one may collect valid and reliable data. The collection of data, even with valid and reliable instruments is, however, only a part of the larger process of research and appraisal. The worth or meaning, in some educational context, that may be assigned to the data obtained is of equal importance. Appraisal is thus a much broader term than measurement since it involves not only the collection and analysis of data but the placing of some value upon it or the reaching of a conclusion regarding its worth. We shall be concerned with both processes in this book.

Objectives as a starting point. All appraisal and research as contrasted with mere measurement or data gathering should be made in the light of objectives, either stated or assumed. Whether the means, processes, and products of a given educational program are adequate will depend upon the objectives sought in each situation. Appraisals made, as they sometimes are, by comparing schools of different countries, states, or nations through the use of conventional criteria are tentative and subject to further or later appraisals in terms of purposes. If we wish to appraise in any substantial way aspects of particular educational programs, we must first know what sorts of products are sought in the particular system and evaluate the extent to which the objectives formulated are being achieved through the projected program.

By starting with objectives one frequently obtains results different from those secured from the application of traditional criteria. Sometimes a very small school system, for example, which cannot meet many generally accepted criteria may yet do a good job in the

[1] A. S. Barr and others, "A Symposium on the Classification of Educational Research," *Journal of Educational Research*, 1931, 23:353–382; 1931, 21:1–22.

[2] Carter V. Good, A. S. Barr, and Douglas E. Scates, *Methodology of Educational Research* (New York: D. Appleton-Century Co., 1936).

light of its objectives. It may also happen that schools possessing many of the characteristics associated with modern schools may still rate low in achieving their professed objectives as well as professionally accepted ones. Even when one is concerned not with over-all evaluations but with evaluations of particular aspects of such educational programs as the teaching personnel, curricula, or administrative organization, effectiveness must be considered from the point of view of purposes and objectives as well as more immediate criteria.

Other reference points in evaluation. There are, besides objectives, other points of view from which evaluations need to be made, as for example, from those of persons, principles, and situations. In making generalizations about persons we must not forget that individuals differ. The research worker in the physical and biological sciences usually works with homogeneous substances. This is not the case in education, where individuals may vary in many respects. Accordingly, our generalizations may vary from group to group, and we must indicate the sort of population for which we expect our generalizations to apply.

Similarly, the conditions under which individuals work vary, and must be controlled or taken into consideration in interpreting findings. Whether we control conditions as in experimental research, or limit our studies to specific types of situation as we may do, we should be careful not to neglect the situational orientation. Purposes, persons, and conditions all provide important points of view from which all research should be oriented.

Every research worker is guided by many principles that he holds to be true. Under certain conditions these provide criteria both for the over-all orientation of research and evaluation, and for the orientation of specific undertakings. Principles provide the general frame of reference for all appraisal and the criteria that one might employ in evaluating particular aspects of the process under investigation.

Wholes versus parts. Research and evaluation usually involve the consideration of both wholes and parts; the one we shall call the telescopic approach and the other the miscroscopic. In the telescopic approach we consider a phenomenon as a whole, its unstructured outline rather than its component elements. The microscopic approach leads to a consideration of the aspects of things, to isolating and separating phenomena into various component parts, and to the study of these individually. It is through microscopic analysis that the research worker detects and segregates certain items for intensive study.

In appraisal and research we use each of these approaches singly

and in combination, and the way in which we use them determines the effectiveness of our investigations. Usually we like to obtain first a general impression. We do this in evaluating persons as candidates for positions, in evaluating educational programs that have been carefully planned, and institutions that we visit. After this initial impression, we may desire more detailed information. We may then analyze more minutely and try to determine strengths and weaknesses, and limiting and facilitating conditions. After we have studied individuals, programs, and institutions in detail, we may then synthesize our findings and ascertain relationships in the more inclusive whole.

The student in studying problem situations will usually follow consciously or unconsciously some such approach as that outlined above going from wholes to parts to wholes.

In some situations the gross total effect is more important than any detailed analysis that may be made of it. For example, the performance of a musician is often judged according to beauty and tone rather than with reference to the mechanical or technical skill exhibited in his performance. However, the latter is the focus of attention under certain circumstances. Some musicians appear to possess the ability to produce an indescribably pleasing effect in their performance whereas others although possessing mechanical skill leave their audience with little lasting impression. In such a case we might say that one musician seems to express a spirit—the over all effect of which is highly effective. The other is recognized for his mechanical ability, but the parts have not been integrated into an effective pattern. In one case we have a gross total impression; in the other we are conscious of mechanical proficiency.

The novice is usually compelled to judge situations or characters by gross estimates. This is only an initial step to be followed by more precise methods of evaluation. He is inclined to make broad generalizations without adequate data and logical checks. He, for example, may conclude without careful analysis that praise as an incentive to learn is superior to reproof in a particular learning-teaching situation or in general. The trained research worker, on the other hand, is interested in determining the precise conditions under which praise may be superior to reproof. His controls and measurements will be carefully applied. As a result he may state that praise is superior to reproof with a particular group of children, in a certain grade, when praise is administered in a certain way, by a certain teacher, and when progress is measured by a certain test. He may also conclude that reproof will be more effective than praise under certain conditions.

Cross-sectional versus longitudinal studies. Another sort of

orientation essential in all research arises from the fact that programs, institutions, and behaviors occur in a continuum. Something has preceded and presumably something will follow. It is impossible to ascertain, for example, whether a particular act is good, effective, or appropriate unless one knows what has preceded and what may follow it. The issue in cross-sectional versus longitudinal research techniques arises principally out of uncertainty about the amounts of a continuum that one needs to consider in answering particular questions. There is no one answer. Sometimes one will desire very detailed developmental-historical studies of the problem under investigation in order to secure the background essential to an understanding of some particular situation. At other times one may be concerned with only the concomitant occurrences and the immediately preceding antecedents. The research worker should give careful consideration to these matters in each problematic situation.

The necessity for comprehensiveness. Need for comprehensiveness becomes particularly acute when we as teachers are confronted at the end of a course with the task of evaluating a student's achievement. At this time, perhaps more than at any other during a course, we are conscious of the need for obtaining evidence of accomplishment that will make possible an adequate summary of the efficiency of particular students. In addition to performance, we may wish to take into account the student's attitude, his effort, and other factors that may bear upon his achievement. This problem includes not only the choice of pertinent information but the assignment of weights to the different circumstances and outcomes.

One of the most important requirements for appraisal and research is that all of the essential elements in a problem chosen for study be taken into account in making the inquiry. It is necessary to consider first one element and then another in a total situation in order that appraisal may be based on the many contributing factors that may be present in particular situations. Comprehensiveness suggests detailed coverage, with stress upon care and judgment in the choice of factors to be considered. Unless all of the important factors are taken into account, we are not in a position to make conclusive generalizations.

The need for qualitative as well as quantitative data. A sound research and appraisal program demands that we use both qualitative and quantitative data. Qualitative data merely indicate the presence or absence of acts, components, and aspects of things whereas quantitative data indicate their amounts. Much of the material needed for comprehensive appraisals requires the use of

data which have not as yet been reduced to a quantitative basis. Among qualitative types of data are those found in behavior descriptions, and various kinds of census compilation. We may also include introspective records of subjects participating in experimental investigations, write-in reports of respondents to questionnaires, verbal reports of individual experiences, and case studies.

In time it may be expected that many qualitative types of data in education will be reduced to quantitative data. It is not too much to hope that in due time most of the so called intangibles, such as interests, attitudes, appreciation, loyalties, and beliefs, will be quantified. Promising beginnings have already been made. It is the complex and not readily observed traits and qualities that are frequently the most important and most difficult to quantify.

Careful planning required. Sound methods of appraisal and research require careful planning, not only as a means of ensuring accurate results but of making it possible for others to repeat an investigation for the purpose of corroboration or refutation.

After a problem has been defined and delimited, the materials, method, and scope of the investigation should be outlined and described. In an investigation requiring use of a quantitative method, statements should be made of the number and kind of subjects used, the instruments of measurement employed, and any other relevant information. Description or explanation of procedure may require considerable detail. The description should be sufficiently explicit to enable anyone reading the study to comprehend its rationale as well as the procedure used, in order that the results may be readily checked, or compared with those of other similar studies, or be verified by further investigation.

The need for accurate instruments. The value of appraisal and research varies with the extent to which observations are accurate. Observations for scientific purposes require objectification and quantification, thus making possible comparisons. We must have adequate observation and measurement in order to compare accurately the performance of any particular individual with that of other individuals; of the performance of an individual at one time with that of the same individual at other times in respect to some quality or behavior. Valid and reliable data are secured only in situations in which the subject's ability to perform a task is measured under standardized conditions and when his responses may be checked according to predetermined or accepted criteria. The uses to be made of the data determine the degree of accuracy required. We wish our data-gathering devices refined to the extent that they provide the kinds of data desired.

Field versus laboratory studies. The laboratory has been regarded as ideal because of the possibility for providing situations that make for effective control of conditions—where changes can be introduced and where variables can operate under direct control. Because conditions of control are not possible to the extent that exists within the laboratory, we have tended to question the accuracy of the field study in education. Yet in the last analysis the field study must constitute our major kind of educational research.

Even though we may not attain the same degree of accuracy in field researches as that which prevails in the laboratory, we can exercise a large measure of control through carefully constructed or selected data-gathering devices, through well chosen sampling techniques, through the use of appropriate statistical devices, and through correct interpretation of results. The field study possesses the advantage of being more lifelike and less artificial.

The applicability of research findings. When we undertake to chart a course of action in the schools, we wish to make sure that every important factor or influence is taken into account so that the program adopted will not be beset by inhibiting factors. We wish our course of action to be such that all our activities may contribute to the attainment of the ends sought. If our objectives have led in one direction and our methods in another, we reshape our policies and methods in such a way as to effect harmonious operation. Planning an educational program, like strategic planning, implies that many circumstances which may be foreseen are considered in shaping policy and in developing methods that will assist in attaining objectives.

After objectives have been formulated and our goals of accomplishment determined, our next problem is to select or adapt means and methods that will serve economically and efficiently in achieving the objectives chosen. Here we may draw upon the findings of research as well as upon informal observation and experience. Certain principles and generalizations, based upon research findings broad in scope and implications, may serve to guide us in our efforts. By drawing upon the results of carefully conducted investigations in the laboratory and the school, we are better able to chart the direction of progress.

The net result is that even though we may have available a body of principles and generalizations gleaned from educational and psychological research, our present responsibility is to approach realistically the problems appearing in the educational program as planned. Skill is needed in adapting the best that research has produced when dealing with educational problems.

There can be no indiscriminate application of research knowledge nor recognition of any ideal practice. Instead we begin with the problems present with all their variables and outline an educational program that will be most effective for a particular situation. During the planning stage and after our plans have been completed we attempt to determine whether our program is in accord with principles and methods that have been derived through research.

PLAN OF BOOK

This book has been planned primarily to discuss methods of appraisal and research. It is concerned with *what* to evaluate only to the extent that the *what* to evaluate influences the selection of methods to be used. It deals largely with problems of *how* to do the job with special emphasis on methods of solving problems in lifelike school situations.

Organization of chapters. The book is divided into certain parts or divisions which are not formally labeled. The first part consists of Chapter II, *Defining Educational Outcomes,* Chapter III, *Quantification of Educational Data,* and Chapter IV, *Criteria of Measuring Instruments.* These chapters establish a basis for the development or choice of dependable data-gathering devices.

If evaluations are to be made from the point of view of objectives, the immediate as well as the broad outcomes of education must be reduced to observable behavior. Because many of the data with which we are concerned in education are highly abstract, operational definitions are essential as a foundation for all research and appraisal. We need only to refer to a few important outcomes of the school, such as interests, attitudes, and ideals, to indicate the intangible nature of many educational goals. The task is to express such outcomes in terms of behavior so that proper limitations may be placed on problems and procedures and so that what to observe in evaluation may become clear.

The accuracy of our evaluations will depend upon the extent to which we are able to quantify aspects of persons, processes, and situations. Chapter III, *Quantification of Educational Data,* outlines the principles that one needs to keep in mind in deriving numerical values for various kinds of data. The chapter deals primarily with the problems that arise in objectifying data-gathering devices that may already be available or in constructing some prior to undertaking a quantitative study.

Following the chapter on the *Quantification of Educational*

Data, the discussion turns to *Criteria of Measuring Instruments.* Two problems arise here: (1) how to evaluate on the basis of accepted criteria instruments that may be available; and (2) how to construct instruments skillfully, if need be, in the light of these criteria. Our problem is frequently not simply one of choosing available instruments but rather of constructing an instrument for a particular purpose or of adapting an existing instrument to the requirements of a study. What should we know about an instrument of research prior to using it in an investigation? What criteria should be recognized in constructing an instrument for a particular purpose? Success of our effort to improve research procedures will depend upon the extent to which we can construct and validate our instruments so that they are sufficiently accurate to sharpen and refine our powers of observation.

The purpose of the second part of the book is to describe some of the nonmathematical and less complex statistical methods of research and appraisal. The division begins with a chapter on *The Description and Appraisal of Status.* This is followed by a chapter on *The Sampling Survey;* and finally by one on *Search for Interrelationships.*

The necessity for describing and appraising status frequently arises in education. Although description and appraisal are generally made without the use of controls this type of evaluation to be most helpful must be better oriented with reference to purposes, persons, and conditions. The criteria employed in evaluations need careful validation, and norms and standards need to be better defined.

By using *descriptive methods* certain valuable kinds of data are secured, and important problems may be solved. These methods, however, frequently serve merely as an introduction to other problems that require further treatment. Chapter VI, *The Sampling Survey,* introduces the student to the important principles of sampling. Here a foundation is provided for the use of quantitative methods of problem solving to be discussed later in the book. This type of study has been extensively used and greatly refined during recent years. In Chapter VII, *Search for Interrelationships,* effort is made to integrate the nonexperimental approaches to the problem of relationship. These approaches include *case studies, comparative studies,* and *historical studies.* Data required here may be variously classified as quantitative or qualitative; subjective or objective; based on current experience or based on past experience. In some instances we are concerned primarily with quantitative data; in other instances the accumulated writings and

researches of others are drawn upon. Here as elsewhere *we are interested primarily in what the investigator is trying to achieve and secondarily with the kinds of data that he may use.*

The succeeding three chapters, *Experimental Design, The Problem of Prediction,* and *Correlational Analysis,* introduce the student to problems that require increasing use of objective instruments for gathering data and making provision for statistical interpretation of results.

The experimental method is especially important because of the possibility that it affords for evaluating hypotheses. The experimenter is not only able to control the phenomena under observation but can produce important elements when he desires. He is able to produce conditions under which certain responses may be elicited. He may modify these conditions and observe different results under varying conditions, repeat former experiments under similar conditions, and make comparative analyses of results. Experimentation enables the investigator to enhance his powers of observation and at the same time modify control over phenomena.

The characteristic which distinguishes the experimental from other methods is the controlled application of *experimental factors.* Controlled group experiments possess two characteristics: (1) there are one or more experimental groups; and (2) these groups are subjected to experimental factors under controlled conditions. Many extraneous factors, including age of pupils, personality of teachers, and size of classes, may influence results of an experiment. These factors may be controlled by modern methods of experimental design and by certain statistical procedures. Experimentation enables the investigator to determine the influence of experimental factors by measuring the performance of individuals or groups, before and after experimental factors have been applied.

The final chapter of the book, *Complex Developmental Studies,* is an effort to provide both a synthesis of the preceding methods of research and appraisal and to show that most successful results can be achieved only by a multivaried attack upon educational problems. It is this many sided approach that makes it possible to reach conclusions that reflect a true picture of conditions as experienced by field workers.

CHAPTER II

Defining Educational Outcomes

The initial task of the evaluator is to analyze the meaning of words that are used to describe educational objectives. The evaluator cannot cope with educational outcomes described in such terms as "proficiency in the art of living" or "ability to lead a good life." Such aims must be expressed in operational terms. What does a person do when he has achieved "proficiency in the art of living"? Is it possible for competent critics to differentiate aspects of a person's behavior that would indicate whether he possesses such proficiency? Is there any likelihood of agreement upon crucial manifestations of such behavior? Only if we know how behavior is affected when attainment of such broad outcomes has occurred can we evaluate the degree of their attainment. It is important that the educationalist cultivate the habit of defining operationally any qualities which he may wish to evaluate.

The thesis that "whatever exists at all, exists in some amount" obviously must be interpreted as including "constructs," which have form only in the human intellect. A construct is a single term denoting an intellectual synthesis of various ideas, such a synthesis being treated as though it existed in reality.

When a term becomes stereotyped, looseness and vagueness in usage are inevitable. Before evaluation can be made of the extent to which individuals possess some quality, there must be substantial agreement upon what constitutes its definitive characteristics. Agreement on such definitive characteristics is possible only when general or abstract terms are analyzed in terms of specific behavior. Otherwise there could be no high degree of accuracy in the communication of ideas.

THE SEMANTICS PROBLEM

The semantics of the term "intelligence" is confusing because of uncertainty concerning the mental traits supposedly measured by intelligence tests. Some writers avoid the difficulty of defining this term by referring to it as "something that intelligence tests purport to measure." Others maintain that each intelligence test constitutes its author's definition of intelligence.

During recent years it has become increasingly evident that intelligence is not a single trait but should be regarded as a composite of abilities. Through use of techniques of factor analysis, intelligence has been differentiated into a number of primary abilities, such as, numerical ability, word fluency, visualization of space, memory for words, names and numbers, perceptual speed, and verbal reasoning. Validity of this differentiation has been challenged; but the attempt to analyze intelligence illustrates a type of approach believed to be essential in attacking the problem of meaning, namely, that of reducing intelligence to significant components before proposing examples of behavior operationally [1] indicative of it.

Statistical techniques of factor analysis are currently applied to problems of this type in an attempt to break down unwieldy general characteristics into relatively distinctive "factors." Such factors, when isolated, can be realistically translated into terms of specific behavior and thus be observed for purposes of evaluation.

Examples of broad outcomes. It will be helpful to consider a summarized formulation of broad educational outcomes: [2]

Objectives of Self-realization:

These objectives refer to zeal for learning; ability to speak, read, and write the mother tongue effectively; ability to solve problems of counting and calculation; skill in listening and observing; understanding of basic facts of health for self, dependents, and community; development of interests in physical and mental pastimes; appreciation of beauty; and ability to give responsible self-direction to one's life.

Objectives of Human Relationship:

These objectives refer to respect for human relationships; enjoyment of a rich, sincere, and varied social life, ability to work and play with

[1] An "operational" definition is one which states the action which occurs or describes the use of something. For example, Thorndike's C.A.V.D. is a good illustration of an operational definition of intelligence.

[2] N. E. A., Educational Policies Commission, *Policies for Education in American Democracy*. Washington, D. C., N. E. A., Educational Policies Commission, 1946.

others; observance of amenities of social behavior; appreciation of the family as a social institution and of family ideals; skill in home-making; and maintenance of democratic family relationships.

Objectives of Economic Efficiency:

The objectives refer to satisfaction in good workmanship, understanding of requirements and opportunities for various jobs; efficiency and desire for improvement in chosen vocation; appreciation of social value of one's work; planning the economics of one's life, with special reference to standards for guiding expenditures, buying skillfully, and protecting one's interests.

Objectives of Civic Responsibility:

These objectives relate to sensitivity to disparities of human circumstances and disposition to correct unsatisfactory conditions; understanding of social structures and processes; defense against propaganda; respect for differences of opinion; regard for the nation's resources; regard for scientific advance according to its contribution to a general welfare; co-operation as a member of a world community; respect for law; economic literacy; acceptance of civic duties; and devotion with unswerving loyalty to democratic ideals.

The major function of such educational outcomes is to give direction to the formulation of relatively specific objectives. Almost every aim that teachers are now being urged to seek may be evaluated upon the basis of one or more such items as those included in the list above.

Outcomes of education have little meaning until we are able to learn precisely what a person does differently from what he did before he attempted to reach certain goals. When we obtain an adequate sampling of the examples of behavior characteristic of an educational outcome, we may then observe the extent to which a given individual "behaves" in accordance with the criterion of performance. It is possible to infer the effect of educational influence from the nature and amount of such performance.

Evaluation of the extent to which education is effectively reaching its broad outcomes is difficult because of the variety and complexity of characteristic behavior involved. The extent to which the more comprehensive outcomes have been attained through the efforts of the educational program is too broad to be appraised by the school itself. The ultimate validation and appraisal of such outcomes is made by society as the individual functions in the social group of which he is a member. Subject to such limitations, however, we may evaluate definable aspects of such outcomes even while the individual is being subjected to educational influences.

Analysis of such aspects is essential, in order to bring specific behavior within the range of evaluation.

To clarify this notion further, we may select for consideration competency in the basic language skills, an aim under "Objectives of Self-realization." This aim must first be brought within the range of tangible meaning. Countless questions may be asked concerning the scope of this item: "What language?" (The answer is presumably the English language.) "What skills?" (Presumably all skills pertinent to the use of language: reading, speaking, writing, spelling, punctuation, grammar.) And "What is basic?" (Expert opinion would undoubtedly be needed to define the scope of "basic.")

After we have defined the area in which evaluation is to be made, our next step is to turn for evidence of attainment to the specific acts which individuals perform. Testing achievement of a third-grade child in spelling English words of third-grade difficulty is a more tangible proposal than that of evaluating a broad outcome. If the instructional objective is formulated as "ability to spell such words as are ordinarily learned upon the third-grade level," the criterion behavior would consist of a practical demonstration that the child can spell such words. If we wish to know how successful a given school has been in teaching for "competency in basic language skills," it is necessary to synthesize all available evidence derived from numerous aspects of the total picture.

EDUCATIONAL OUTCOMES IN TERMS OF BEHAVIOR

The problems with which we are concerned are (1) clarification of the meaning of the objective or characteristic to be considered, and (2) detailed definition in terms of specific behavior. A point of view with which to approach the treatment of these problems is suggested in Tyler's formulation [1] of the second basic assumption of the Evaluation Staff of the Eight-Year Study:

> A second basic assumption was that the kinds of changes in behavior patterns in human beings which the school seeks to bring about are its educational objectives. The fundamental purpose of an education is to effect changes in the behavior of the student, that is, in the way he thinks, and feels, and acts. The aims of any educational program can not well be stated in terms of the content of the program or in terms of the methods and procedures followed by the teachers, for these are only means to other ends. Basically, the goals of education repre-

[1] E. R. Smith, R. W. Tyler, and others, *Appraising and Recording Student Progress* (Harpers, 1942, p. 11).

sent these changes in human beings which we can hope to bring about through education. The kinds of ideas which we expect students to get and use, the kinds of skills which we hope they will develop, the techniques of thinking which we hope they will acquire, the ways in which we hope they will learn to react to esthetic experiences—these are illustrations of educational objectives.

The purpose of this chapter is to illustrate the problem of defining educational outcomes by analyzing a number of areas in which evaluation is likely to occur. No effort will be made to make the treatment exhaustive. The materials to be presented are simply illustrative of the many situations that might have been analyzed. We shall discuss first some of the simpler and more tangible areas in educational situations. These will be followed by situations which are more intangible and abstract.

MOTOR SKILLS

When analyzing human qualities we are almost always dealing with the combined results of mental and motor learning. All behavior involves some type of motor adjustment manifested in spatial perception or in some overt physical movement. When evaluation is concerned with extensive amounts of motor learning, it is easy to overlook the ultimate dependence of motor performance upon mental learning.

Ability to perceive spatial relationships involves a complex coordination of mental and motor activity. Spatial perception is involved when an individual visually explores a route through a maze before attempting to trace his way through it.

Ability in spatial perception may be demonstrated without recourse to use of symbolic language. It may occur with little "inner speech" in the form of conventional symbolic language. It is demonstrated without great use of language skills on an intelligence test by exhibiting simple physical activity, such as that required in checking correct or incorrect forms or in drawing lines. Nonverbal elements are commonly included in intelligence tests not only because they are associated with intelligence but also because spatial perception may be evaluated without verbal expression. The Army Beta test is a celebrated example of an intelligence test consisting of nonverbal elements which were designed to measure mental ability of persons who were handicapped by lack of education.

Tests of physical performance are administered to measure the general development of individuals, usually children, on the basis

of abilities related to neuro-muscular co-ordination. Many compilations have been made of acts which children are able to perform at various ages. A number of simple acts are performed without elaborate testing equipment and interpreted as evidence of general development. Such acts of the child may be to sit erect on the floor or to grasp something with one or both hands. During a child's early years various types of motor ability are indicative of mental maturity. Intelligence tests for young children are almost always based predominantly upon motor performance. Such behavior may be directly observed and differs from that manifested in the form of written language, as in pencil-and-paper tests, in which only symbolic behavior is noted.

Performance tests are also administered to predict ability to perform physical tasks peculiar to certain vocations. Many occupations require workers who possess well-developed manipulative skills. For purposes of selecting such individuals, performance tests may require the individual to use mental and motor skills similar to those supposed to be operative in an occupational situation.

During World War II considerable study was made of specific motor and mental abilities necessary for certain technical duties in aircraft operation and performance tests that would effectively select the most desirable candidates for special training. Many performance tests require such activities as tracing a maze, fitting objects of varying shapes and sizes into recesses in a form board, detecting omission of details in pictures, counting variously piled cubes, and identifying similar forms. All these are among the techniques of causing the individual to reveal different aspects of spatial perception and of mental and motor coordination. Other instruments include tests of reaction time, agility and strength, steadiness, finger dexterity, and ability to assemble simple mechanical devices.

Typewriting skills. The role of behavior as a basis for evaluation may be clarified by considering certain instructional aims sought in teaching typewriting. The following objectives are appropriate for consideration: [1]

1. To develop manipulative skills and to learn to use the parts of the machine expeditiously.
2. To review English grammar on a functional basis and to learn those elements of usage peculiar to typewritten or printed materials.

[1] E. Popham and I. Place, "Aims of college typewriting," *Journal of Business Education*, 1947, *27:* 17–18.

3. To build good work habits and to evolve orderly procedures for handling routines.
4. To develop an understanding of and appreciation for the typewriter as a writing instrument, so that the operator will care for it properly and see that it gets needed repairs.
5. To assume responsibility for proofreading one's own material.
6. To sustain a typewriting rate which assures production of a reasonable amount of usable work over a considerable period of time.
7. To compose usable copy directly at the typewriter.

An intimate relationship exists among the mental and motor skills described in these objectives. The objectives include types of mental learning necessary for performing many of the duties which might be performed by a secretary. A typist-secretary might, for example, be required to punctuate a dictated letter, to center a title on a page or to compose an appropriate form of tabulation for unfamiliar types of data. Efficient performance in such instances depends not only on the ability to make appropriate physical movements at the typewriter keyboard but to do so in connection with the purposes for which typewriters are used. Desired progress would not be made if it were limited to mastery of the skill of applying the proper fingers to the various areas of the keyboard and manipulating such accessory controls as the space bar, the carriage-return lever, the backspace key, or the tabulator key.

Typical learning standards [1] may include some of the following: stroke 40 words a minute for one-minute periods of practice material at the end of six weeks of instruction; change a typewriter ribbon in one minute; tabulate three columns of unarranged material in ten minutes; center three short lines of unequal length in five minutes; type 40 words a minute for ten minutes, including erasure time and address envelopes correctly at the rate of one every 90 seconds. The criteria of manipulative mastery are empirically defined in amounts and types of product within time limits. Speed is commonly determined by computing the number of five-letter words which may be written in one minute. This approach results in a series of operational definitions of the behavior which the individual displays at different times throughout the period of his training.

Courses in typewriting are conventionally adjusted to the demands made of typists in business and commercial situations. In

[1] E. Popham and I. Place, "Aims of college typewriting," *Journal of Business Education*, 1947, *27:* 15–16.

order to approximate typical vocational situations, materials used for practice are those commonly used in occupations involving typewriting. Although business letters constitute the dominant type of material, performance situations include numbers, tabulation of data, cutting stencils, and activity other than that required in the copying of simple verbal material. Consequently, the individual's performance is analyzed with reference to two forms of behavior: (1) performance manifested in motor action and (2) performance expressed in varying types of typewritten material. Both types of behavior may be observed during performance and in the results of performance. These may be evaluated according to the extent to which the behavior revealed indicates progress toward instructional objectives.

Observation plays an important role in the appraisal of many gains made by pupils in typewriting, especially in connection with work habits, neatness, care of typewriter, and methods of handling material without unnecessary motion. For effective appraisal it is desirable that such goals as "efficient work habits" be reduced to specific activities.

Industrial arts curriculum. Similar analysis may be made of the industrial arts curriculum, in which many motor skills are developed. Objectives of the industrial arts curriculum are adjusted to the acquisition not only of desirable motor skills in specific shop areas such as woodworking or automobile mechanics but of an adequate informational background related to industrial practices. The outcomes of the curriculum may be expressed in the following formulation, which is subject to modification according to grade level, maturity of learner and community interests: [1]

1. Ability to express one's self through planning and constructing projects, using common tools and a variety of construction materials, typical of industry.
2. Discovery of aptitudes and reactions contributing to the maturity of life interests, both of a vocational and an avocational character.
3. Understanding of industry and its products and services, together with their influence in determining patterns of living in modern society.
4. Ability to read and make working drawings for planning and constructing useful projects typical of modern industry.

[1] M. M. Proffitt and others, "The measurement of understanding in industrial arts," Ch. XVI, in *The Measurement of Understanding*. N.S.S.E., 45th Yearbook, Part I, 1946, p. 303 ff. Reprinted by permission.

5. Ability to choose industrial products with reference to design, pleasing color combinations, and durability; and to maintain and service such products.
6. Growth in abilities and attitudes related to mathematics, science, and the language arts, and to work habits, safety practices, and co-operation with others.

Evaluation of the effectiveness of the industrial arts curriculum would necessitate a synthesis of evaluations of achievement made in certain designated areas. Instructional aims are best defined operationally when the frame of reference is relatively homogeneous. In woodworking, for example, such aims are related to a specific type of material and to tools having relatively specific uses.

As in typewriting, skills relating to physical manipulation in woodworking may be observed while behavior is occurring and also in tangible end products. An instructor may observe an individual while he makes cutting lines on a piece of wood; uses a saw in cutting along such lines, smooth surfaces, and edges with proper types of plane and sandpaper; and applies stain and varnish. He may inspect a completed table or cabinet for evidence concerning the workmanship of assembly.

In woodworking, considerable emphasis must be placed on observation of work in progress, since it is during progress that significant evidences of skill are displayed. The most valid evidence that the learner is able to use a plane correctly is exhibited while he uses it. His observed behavior also reveals the extent to which he is able to visualize the completed product, construct its various parts in orderly sequence, and foresee the proper fitting of parts upon assembly. Adequate evaluation in the area of physical performance necessitates detailed formulation of specific operations which the learner should be able to perform. By an analytic approach it is possible to determine areas of achievement in which further improvement may be made and to obtain an accurate picture of achievement in a general area of related motor skills.

Evaluation of motor performance, however, may only partially reveal total achievement in woodworking. In appraising achievement in any industrial arts field, it is important to evaluate the accomplishment of the learner not only in respect to skills which may be displayed in physical action but also in respect to mental learning. Desired types of mental learning are included among the outcomes previously listed for the industrial arts curriculm. Instruction in a specific area may include use and maintenance of tools, ability to plan proper sequence of operations, measurement

of lumber, qualities of different varieties of wood, and vocabulary to describe techniques used in fabricating wood products. Range of information and understandings may not be thoroughly appraised in the individual's performance with woodworking tools. If such achievement in mental learning is included in the intent of objectives formulated, it may be desirable to extend evaluation by using pencil-and-paper tests.

The following item [1] is suggestive of understandings which may not be revealed during the physical activities of woodworking:

Check the statement which constitutes the best answer to each question:

1. Furniture makers prefer to work with mahogany because mahogany
   a. is heavier.
   b. does not require a filler.
   c. is less likely to chip (flake).
   d. is more plentiful.

Objectives in fields of instruction in which physical activities are important are characteristically displayed in a background of intellectual understandings. Students of physical education not only seek improvement in physical performance but discover that the field requires knowledge of health, physical fitness, bodily growth, and social participation. Students of home economics seek proficiency in various manually expressed arts connected with homemaking activities. The field of home economics also includes a wide diversity of intellectual understandings related to management of family life.

Ability in various physical activities is usually best evaluated by noting the physical behavior in which the individual is capable of performing and the product which he can produce. From observation of such behavior we may infer to some extent the presence of intellectual understandings.

Types of observable behavior in the case of directly taught motor skills are similar to those which have been operationally defined by formulated instructional objectives. They are manifested in an extensive amount of overt behavior, primarily because physical movement is directly observable. Such behavior increases the likelihood of more valid appraisals of motor skills than is often the case when evaluation of some mental skill or of some emotional quality is attempted. In typewriting, for example, the learner is visibly engaged in many elements of the type of behavior which

[1] M. M. Proffitt and others, *op. cit.*, p. 306.

is desired. Not only may an end product be exhibited but the process through which it is derived may be evaluated during occurrence.

BASIC MENTAL SKILLS

Language and arithmetic skills are commonly regarded as prerequisite for even minimum attainment of educational outcomes. Both types of skill relate to ability of individuals to communicate. In a sense mathematics involves essentially the ability to use the language of quantification which may be expressed verbally or by means of mathematical symbols. We may say, for example, three plus two equals five or write "$3 + 2 = 5$." Language and arithmetical skills, however, are considered separately because they involve fundamentally different symbols; systematic uses of each class of symbol are different.

Language skills. Definition of the objectives sought in the cultivation of language skills requires analysis of certain relevant general behavior into a number of types of specific behavior. Reading, for example, is a highly complex general behavior. Many of its component specific kinds of behavior must be appraised individually in order to evaluate the extent to which reading ability is being acquired. When the specific objectives essential to the cultivation of reading ability are formulated, account must be taken of the kinds of behavior that different individuals can demonstrate as evidence of improvement.

In learning to read, improvement may be expected in at least four areas:

1) *The individual's knowledge of vocabulary increases.* He comes to know the meanings of more and increasingly difficult words. Additional meanings for words already known in connection with a single meaning are acquired.

2) *Progress occurs in the techniques of usage and spelling.* The individual acquires a foundation of experience with which he may later make comparisons, saying, for example, "This sentence sounds correct" or "This spelling looks right." A foundation is thus laid for ability to recognize good usage and correct spelling.

3) *Gradual increase in perceptual span occurs.* Eye fixations per line of reading material gradually decrease in number and larger word groups are perceived.

4) *Simple inferences become increasingly habitual.* The individual to a greater extent can impart meaning to what he reads.

This improvement may result from practice in reading or from gradual increase in the number and significance of daily experiences.

There is also an increase in speed and accuracy. In most instances speed and accuracy develop simultaneously. The ability to read rapidly is associated with ability to read accurately.

To analyze comprehension as an aspect of reading ability, inquiry is first made as to what individuals do when they are able to read with comprehension. The setting for evaluation is prepared by describing the activities that the individual should become able to demonstrate, listing them as objectives of instruction, and then by devising performance situations. One of the difficulties in evaluating comprehension is that of defining adequately what is meant by comprehension. Comprehension in reading may refer to simple comprehension or to *verbal reasoning and organization.*

An empirical approach to the interpretation of simple comprehension may be made by restricting the performance expected in response to a reading selection. Ability in simple comprehension is demonstrated if a satisfactory number of individuals of a given grade can read a selection and then reproduce in their own words the information, answer questions related to informational content, or recognize truth or falsity in statements related to the content.

Ability in verbal reasoning and organization may be evaluated by requiring several types of performance. Understanding the significance of a given selection of material is demonstrated when pupils can satisfactorily summarize the central theme of a paragraph or state the field of knowledge to which a paragraph is related. Performance may consist in predicting the outcome of certain events partially presented in a reading selection.

Comprehension may also be demonstrated in ability to carry out instructions related to performance of a task. Such a task might be to circle, underscore, or cross out designated letters in a few sentences. In another type of performance, the individual may be required to indicate the sequence in which three or four statements should occur in order that a paragraph possessing unity and coherence may result. Accuracy of perception may be tested by asking whether some apparently unimportant word was observed.

Davis [1] surveyed a large number of reading investigations for

[1] F. B. Davis, "Fundamental factors of comprehension in reading," *Psychometrika,* 1944, 9, 186.

the purpose of discovering abilities believed by investigators to be significant. This survey resulted in the tabulation of several hundred specific skills. These skills were classified in such a way as to include those that appeared to belong in the same categories and thus closely related. The different categories therefore represented relatively distinct abilities. On the basis of this analysis and classification nine groups of reading skills emerged as follows:

1. Knowledge of word meanings.
2. Ability to select appropriate meaning for a word or phrase in the light of its particular contextual setting.
3. Ability to follow the organization of a passage and to identify antecedents and references in it.
4. Ability to select the main thought of a passage.
5. Ability to answer questions that are specifically answered in a passage.
6. Ability to answer questions that are answered in a passage but not in the words in which the question is asked.
7. Ability to draw inferences from a passage about its contents.
8. Ability to recognize the literary devices used in a passage and to determine its tone and mood.
9. Ability to determine a writer's purpose, intent, and point of view, i.e., to draw inferences about a writer.

With these nine categories operationally defined Davis proceeded to construct test items which would measure the abilities designated. The study illustrates several steps which will assist the research worker in initiating a study. First, the author made certain that he was familiar with the vast amount of research and writing available in the field of reading. These materials afforded a basis for analysis and classification of available knowledge. Secondly, he expressed these data operationally—in terms of what pupils should be able to perform as evidence of reading ability. And thirdly, he used these operational definitions as a basis for constructing a measurement instrument in reading.

Mathematical skills. Mathematical skills require three general types of activity: (1) manipulation of numbers on the basis of rule or memory, (2) use of numbers or letters in abstract reasoning, and (3) reasoning involving spatial or geometric relations. Instruction on all grade levels provides practice in problem solving by means

of numbers or letters. These activities become increasingly difficult in the order given. Throughout instruction in mathematics, the student develops increasing mastery in respect to number concepts, use of mathematical symbols, number relationships, and solution of problems through use of numbers. Number concepts deserve brief explanation. The student develops mastery of number concepts as he becomes progressively familiar with writing integers, monetary symbols, fractions, Roman numerals, fractions and decimals, exponents and roots, negative numbers, and abstract numbers.

The four basic arithmetic skills involving numerical manipulation are addition, subtraction, multiplication, and division. These skills are regarded as the fundamental tools for all subsequent study of mathematics. Specific objectives have been formulated with an exceptionally high degree of precision throughout the field of mathematics, particularly at the levels on which the basic skills are first taught.

The four basic manipulative processes constitute operational concepts; and in all four, specific behavior has been precisely outlined. Some of the specific problems relating to addition are combining small numbers, addition of columns, addition of mixed numbers, addition of numbers when used denominately with units of measure, reduction of fractions to a common denominator, addition of fractions and decimals, and the writing of decimals in columns and adding percentages.

Objectives in learning arithmetic are concerned with mastery of processes to a degree of ability at which the operations involved may be correctly performed regardless of the numbers used. In general, the goals sought in the fundamental skills may be outlined with such precision that the instruments of evaluation are usually of high diagnostic value.

Analysis is made upon the basis of whatever levels of difficulty or complexity have been established for a given grade or age level. For example, certain types of addition are not expected before the eighth or ninth grade. Among the simple aspects of addition, the pupil may be expected to demonstrate that he can (1) perform typical examples of addition, (2) solve problems requiring addition, (3) decide whether addition or some other process is required to obtain a correct answer, and (4) decide what classes of object may or may not be added. He may, for example, add 2 books and 3 books, but he may not add 2 wheelbarrows and 4 rakes. Instructional objectives for each grade or age may outline precisely the types of behavior that pupils can be expected to demonstrate

when these objectives have been attained. Instruction toward such objectives consists of practice in each type of behavior up to a degree of difficulty at which relatively few members of a group can perform successfully. For the purpose of measuring achievement, examples and problems typical of those that have been practiced are used as tests.

ABILITY TO APPLY PRINCIPLES

The Evaluation Staff of the Eight-Year Study was confronted with the problem of analyzing certain objectives which had received considerable lip service but which had been neither satisfactorily analyzed nor accurately measured by available tests. Improvement in various aspects of thinking was a desired goal of such instruction. Various operational aspects of this goal were defined for various subjects. One of these goals was the ability to apply principles of science. We shall describe the analysis of this objective and its definition in terms of specific behavior.

In order to clarify this objective, it was first necessary to analyze the behavior involved in making applications and to select principles of science with which teaching and testing would be concerned. Analysis revealed that two operations were involved in the mental process of making an application: (1) the individual examines a situation and tentatively selects an explanation, and (2) he then uses any necessary principles of science and good reasoning in order to justify his tentative solution. His thinking consists of a search for a general rule and his final decision involves consideration of similarities believed to exist between the given situation and other situations in which the general rule applies.

In constructing instruments for measuring progress in the attainment of ability to apply principles, it was agreed that all situations requiring application should be different from situations which had been previously practiced. Otherwise, the test would require recall of a principle already found appropriate to the situation presented, rather than one that would require actual performance of the mental operations involved in selecting a principle that would be applicable. The term "principle" referred to any law, generalization, or understanding of science that might be proposed as a formulated reason for an explanation.

Selection of appropriate principles to be stressed in the fields of chemistry, physics, and biology in the secondary school was undertaken by teachers co-operating with the Evaluation Staff. From the same source items were obtained for problem situations. Each item was required to be (1) new to the student or not generally used as

sample illustrations in classroom discussion or in textbooks, (2) typical of situations commonly encountered in daily life, and (3) capable of explanation by one or more of the selected principles.

From the situations contributed by teachers, selection was made from those which would require of the student one of four different types of response: (1) prediction of what would happen, (2) explanation of what had happened, (3) choice of a proper course of action, or (4) acceptance or rejection of proposed explanation. The test forms ultimately adopted were designed in such a way that the student's response was restricted to explanations proposed as plausible and to principles that he might decide to use. These limitations contributed to objectivity in scoring without unduly limiting opportunity for full expression.

The following problem [1] is reproduced, with modification, from a test developed for measuring ability to apply principles of science.

PROBLEM:

The water supply for a certain big city is obtained from a large lake, and sewage is disposed of in a river flowing from the lake. This river at one time flowed into the lake, but during the glacial period its direction of flow was reversed. Occasionally, during heavy rains in the spring, water from the river backs up into the lake. What should be done to safeguard effectively the health of the people living in this city?

Directions: Choose the conclusion which you believe is most consistent with the facts given above and most reasonable in the light of whatever knowledge you may have, and mark the appropriate space on the Answer Sheet under Problem.

Conclusions:

A. During the spring season the amount of chemicals used in purifying the water should be increased.
B. A permanent system of treating the sewage before it is dumped into the river should be provided.
C. During the spring season water should be taken from the lake at a point some distance from the origin of the river.

Directions: Choose the reasons you would use to explain or support your conclusion and fill in the appropriate spaces on your Answer Sheet.

Reasons:

1. In the light of the fact that bacteria can not survive in salted meat, we may say that they can not survive in chlorinated water.
2. Many bacteria in sewage are not harmful to man.

[1] E. R. Smith, R. W. Tyler, and others, *op. cit.*, p. 91–93.

3. As the number of micro-organisms increases in a given amount of water, the quantity of chlorine necessary to kill the organisms must be increased.
4. Sewage deposited in a lake tends to remain in an area close to the point of entry.

It may be observed that in order to deal successfully with this item the student must actively engage in the types of mental activity involved in making applications of principles of science. The item was framed in accordance with the mental operations regarded as indispensable for its correct solution. Recall of information is inevitably involved in behavior requiring complex mental ability, but recall alone is not sufficient for a correct response to be made. The purpose of the item is to require the student to perform the mental operations attached to the type of complex ability in which progress and improvement are desired.

TRAITS OF PERSONALITY

Personality is revealed in the effects which individuals have upon one another and in their behavior in social situations. Certain individuals, for example, may react to the problems of life with relatively low degrees of confidence and assurance, whereas others are seldom anxious over the likelihood of success in any of their undertakings. In studying personality, we are principally concerned with the individual's outlook upon the demands of social situations. We are critical of differences in traits of personality primarily from an objective point of view. Our aim is to clarify the various socio-emotional tendencies of the individual.

It is conventional practice to consider traits of personality paired as opposites. An individual may be at one extreme of "extroversion" or at the alternate extreme of "introversion." The direction of a trait depends upon which extreme is dominant as judged by competent observers of the individual's behavior. The intensity of a trait refers to the extent to which such behavior places him close to either extreme. Evaluation of traits of personality is a process of determining the direction and intensity of given traits.

In many instruments used to measure traits of personality, effort is made to induce the individual to reveal himself in terms of behavior which he believes is typical of himself. Items such as the following may be presented to him:

1. Do you feel disturbed toward the end of a dinner over the possibility of being called upon to make a few remarks?
   Yes. No.

2. Do you cheerfully accept criticism concerning the way in which you are performing a task?
   Yes. No.

The implication of these items is that a "yes" or "no" answer might be indicative of the extent to which an individual tended toward extroversion or introversion. The purpose of such an item is to determine whether the behavior of the individual resembles that of persons who tend to avoid revealing themselves or that of persons who tend to regard daily life objectively.

The goal in developing a science of personality is to devise procedures for describing individuals. The purpose is to use as few trait names as possible yet at the same time to characterize a wide range of behavior. It is convenient for some purposes to classify individuals as "leaders" or "followers." It is evident, however, that only a small area of socio-emotional behavior is directly related to "leader" and "follower" categories. To characterize individuals with respect to only one such trait would fail to afford much information of value concerning a large group of individuals who may not be validly called either "leaders" or "followers." There may be occasions however when the use of several traits may be necessary to increase assurance that we have overlooked no important aspects of personality.

After the characteristics of behavior relevant to a given situation have been determined decisions may be made about the traits concerned. The socio-emotional significance of a large number of specific acts which individuals perform or avoid performing are considered. These acts are selected with special reference to their significance as trait-indicators, and limited to the actual clues concerning the nature of an individual's inner qualities, that is, all acts which the possessor of a trait might perform but which he would not perform if the trait were absent. It is the performance of acts which distinguishes the individual in respect to some quality that is important. It might not be significant to discover that an individual has failed to fill his fountain pen; but the effect of habit might be observed in his act of filling it every morning; such an act might be related to some defined trait of personality.

The acts chosen should be habitual reflections of the attitudes or interests. It is the habitual nature of an act that is significant and this is frequently expressed by using such words as "usually," "often," "ever," or "generally" in the statements or questions describing the act. An example of a typical item is: "Do you usually feel tired when you get up in the morning?" Very often the predi-

cate used in the item implies a continuity of action during a period of time as, for example: "Do you believe that typical modern literature is unwholesome?" The idea of continuity is implied by the fact that time is a factor involved in the mental process of believing. Habit may also be implied by the form of the verb as, for example: "Do you take long walks alone?" or "Do you keep a diary?" Even when qualified by such a word as "ever" a statement may make reference to a habit as, for example: "Have you ever stolen anything?"

Personality is significantly linked with the individual's patterns of action which are essentially tendencies to act in a consistent manner. A major purpose in studying personality is to obtain information which affords a basis for making predictions. The probability of predicting the nature of an individual's course of action depends in part upon the extent to which specific acts set forth in an inventory of personality are indicative of a trait.

INTERESTS, APPRECIATIONS, AND ATTITUDES

Interests. Interests are characterized by voluntary self-identification with some activity. Upon his own volition a child develops interests in books, newspapers, radio programs, movies, personal hobbies, and in activities related to various fields of study. At an early age he may indicate preferences for certain vocations, even though such preferences may not become stabilized for many years. The value of a pupil's interests for the general aim of education lies primarily in his motivation. This must be discovered and appropriately related to school activities because of the emotional satisfactions involved. A wide variety of wholesome interests tend to affect school activities favorably.

Interests are, for education, both a means and an end. Interests can be used to motivate school work and to make various types of subject matter meaningful. An important outcome of education should be the extension and improvement of existing interests. In evaluation, therefore, we are concerned in determining not only the nature and quantity of interests that an individual possesses but also the extent to which they are significant as (1) contributors to educational outcomes and (2) educational outcomes in their own right.

Surveys of interests may be made by observing what pupils select as voluntary activity or by using such various devices as questionnaires or check-lists to enable them to indicate the kind and quantity of their interests. Surveys may also prove of value when

undertaken within limited areas. For example, it may be indicative of interests to know the titles of books read during a given period or the number of pages read within a given book. We can not evaluate an interest, however, until certain criteria are established. Some interests may be characteristic of the individual's stage of development and yield to other interests as the individual increases in age. As educational outcomes, the individual's interests that are most essential are those that have value for him in various aspects of living, such as his home, economic, civic, and recreational activities.

The Evaluation Staff of the Eight-Year Study[1] selected for evaluation interests in reading. Four aspects of reading interests were employed. Interest in reading was considered from the standpoint of *abundance, variety, selectivity*, and *maturity*.

Abundance of reading refers primarily to quantity of material read. It may be evaluated by obtaining various evidences of pupil behavior in respect to the number of books read as reported by pupils or as determined from library circulation data. Frequency of book reading is assumed to have a bearing on the intensity of the reading interests and perhaps to a lesser extent on its quality.

Variety of reading is related to the extent to which the pupil reads different types of material, for example, the amount of fiction or nonfiction. Reading interests may be explored not only in relation to books but also in respect to newspapers and magazines; in the latter with a view to determining how wide a range of periodicals is habitually read.

Evaluation may be concerned with *selectivity* of reading. A high degree of selectivity is suggested when individuals are found to concentrate on one or more areas of reading—as, for example, on radio construction, history of music, or habits of animals. Pupils may read abundantly and yet do so with little discrimination as to types of material. Increase in selectivity presumably indicates an increase in the extent to which reading has acquired additional personal value to the individual, whereas abundance may simply indicate that reading is a favorite means of using free time.

Maturity of reading interests refers to the extent to which books chosen are difficult or require considerable insight. The point of reference may be the types of reading commonly engaged in by adults and by growing children at different age or grade levels. Evaluation of *maturity* by means of this criterion necessitates the appraisal of different types of books with respect to the maturity of the individuals who ordinarily read such types. A pupil's reading

[1] E. R. Smith, R. W. Tyler, and others, *op. cit.*

may be considered typical if it includes many titles of books appropriate for his maturity level.

In general, evaluation of interests is possible on the basis of whatever qualities are revealed by consistent behavior. We are interested in how such qualities of an individual's interests compare with those of other individuals or how they appear as a profile of interests in various areas. Even after we have collected extensive data concerning the interest of an individual in terms of *abundance, variety, selectivity,* and *maturity,* it is necessary to interpret the significance of these criteria for the individual. We do this by determining whether they are consistent with his total pattern of interests, whether his total pattern is reasonably well integrated for his age and maturity, whether they can be made to contribute to meaningfulness of school subjects, and whether they should be encouraged or discouraged as educational outcomes.

Appreciations. To distinguish between an appreciation and an interest requires careful analysis. An interest is usually restricted to some activity that the individual likes or in which he may voluntarily engage, even though he may possess little exact knowledge or understanding of it. One of the distinctive characteristics of an appreciation is insight into the value of the object appreciated. Etymologically, appreciation is associated with the notion of price or value. One's appreciation of grand opera may be due to the fact that he has developed sensitivity to the merit of an operatic production.

A person who has developed appreciation of good music is usually one who is not only emotionally responsive to music but who has acquired a background of training or experience which serves as a basis for his appreciation. One who appreciates good music may be prevented by lack of opportunity from becoming creative in music. Appreciation involves capacity to make an emotional response and does not necessarily depend upon expert knowledge or skill. An individual may know the characteristics of various types of art, be familiar with many conventions for use of color, and shadow, but find himself in disagreement with experts who possess a much more highly developed gift of appreciation.

An appreciation must be analyzed with reference to an individual's behavior within the scope of some specific area. Evidence concerning the nature and force of an appreciation in reading may be obtained by observing such aspects of his behavior as the following:

1) *Does the individual derive sufficient genuine enjoyment from certain types of reading to the extent that he might wish to repeat his experiences?*

2) *Is the individual stimulated to read more material of similar nature?* Does he use enjoyed reading as a criterion for selecting further reading material?

3) *Does the individual engage in self-identification with any of the characters of a story or can he imagine himself being in places described?*

4) *Does he feel that the material read has been of personal value to him, has applied to any of his problems, or has influenced his thinking?*

5) *Does the individual feel definitely that the material read possesses merit?* Does he discover himself evaluating the material read?

6) *Does the individual wish that he himself might have written the material read or that he might write something similar?*

The Evaluation Staff [1] of the Eight-Year Study sought to evaluate art. They tested the assumption that ability to detect significant similarities and dissimilarities in art objects is a satisfactory criterion for an objective evaluation of appreciation in art. The technique consisted in preparing "picture sheets" bearing copies in color of many more or less well-known paintings representative of different periods of art. The task was to match as many pairs of pictures as possible on the basis of style, use of color, composition, or mood. Specific instructions were given not to select pairs on the basis of similarity of subject matter. Behavior was appraised according to consensus of expert opinion.

Although the individual's behavior constitutes the only readily accessible evidence of educational outcomes, utmost care must be used when one identifies and measures appreciation. In appreciation it is often probable that aspects of behavior might have been considered other than those chosen for analysis. That which constitutes crucial behavior depends upon the initial definition of the trait under consideration. Many of the less tangible outcomes of education present difficulty in evaluation because of the vagueness of the terms used.

Attitudes. An attitude may be defined as a learned emotional response set for or against something. Its directional aspects are usually more conspicuous than is true of an interest or an appreciation. Attempts are seldom made to explore the negative aspects of an interest or an appreciation. Attitude-objects may be extremely general, as in the case of an attitude relating to conservatism or liberalism, or extremely specific as in the case of one's attitude concerning vaccination.

Evaluation of an individual's attitude is concerned primarily

[1] E. R. Smith, R. W. Tyler, and others, *op. cit.*

with determing the *direction* and *intensity* of his feeling for or against some belief. This belief should first be clearly identified. The direction of a given attitude is not directly involved in the evaluation; the moral implications of an attitude constitutes a separate problem. The purpose of evaluation is to determine the extent of change, especially when producing a change is one of the goals of instruction. A goal sought in a course in civics may be to establish attitudes related to fulfillment of obligations of citizenship. At various points during a course, the effects of instruction upon attitudes may be evaluated in order to modify procedures in producing change.

Attitudes are revealed in a more adequate degree when an individual's behavior is overt. It is often difficult, however, to make a direct observation of behavior. In evaluation of attitudes, opinions expressed by the individual himself must usually be accepted. He may also be rated by some group of persons as to the probability of certain behavior. Changes in intensity of an attitude may reach a neutral point between favor and disfavor. If we obtain no evidence concerning its direction or intensity, it is implied that either no attitude exists or its effects are not measurable. An individual, for example, may hold an indifferent or an extremely mild attitude toward the desirability of foreign missions. If he is asked to express an opinion in connection with foreign missions, he may be influenced by attitudes that only remotely relate to the opinion expressed. He may conceive of the problem not as a religious one but as one related to certain attitudes that he holds in favor of national self-sufficiency as opposed to international-mindedness. The fact that an individual's opinions will often reflect his broad attitudes concerning specific issues may constitute a technique of evaluation. An individual's behavior is usually influenced by attitudes which possess a fairly high degree of consistency during a period of time.

One approach to the evaluation of attitudes has been through the use of questionnaires. Lists of operational statements are assembled, and the individual whose attitude is being evaluated is asked to record his probable reaction toward the various types of behavior proposed. Each statement is pointed to one extreme or the other to the attitude that is being appraised. A wide variety of situations may be presented. It is possible in many cases to prevent the individual from detecting the nature of the attitude which is being evaluated. A "yes" or "no" response, however, may afford little opportunity for the individual to record the intensity of his beliefs. In order to increase the discrimination of a questionnaire, opportunity is often provided for the individual to qualify his re-

sponses. This would be possible if such items as the following were used:

R+ R ? W W+	Medical service should be supported by our government in order that all persons regardless of their circumstances may enjoy necessary medical care.
R+ R ? W W+	Private industry is usually more efficient than public ownership and operation.

Each of the foregoing statements is relevant to behavior of the individual, since he would presumably act consistently with his response if opportunity were available. Such items might be used in order to determine whether the individual is conservative or liberal. The behavior suggestive of a point of view need not propose a specific course of action, since the purpose of the item is to obtain an expression of approval or disapproval. The implication in this technique is that the individual may envisage the future action of an approved situation. The questionnaire is relatively easy to construct, but is ill adapted for scaling accurately the intensity of. an attitude.

The scale-type instrument attempts to minimize this difficulty by defining the attitude in terms of a single attitude-object. All items, therefore, may be constructed with gradations of favor or disfavor, thus enabling the individual to register the extent of his attitude. Each item is evaluated by consensus of opinion as to its bearing upon the attitude and is assigned a score value. Such a scale may relate to "Attitude toward Public Ownership and Operation of Basic Industries." The following items present points of view varying in intensity and direction:

Write a check mark (√) if you agree with the statement. Write a cross (×) *if you disagree with the statement.*

( ) 1. The government has a definite responsibility for enabling the poor man to purchase essential commodities at a low price.

( ) 2. The present strength of American industry is a result of competition among producers to produce the best product at the lowest price.

( ) 3. Competition between food producers results in added costs due to advertising.

( ) 4. Advertising results in greater consumption of advertised goods and consequently greater production at lower unit cost.

Several other types of scale have been constructed. One of these makes possible the measurement of intensity by use of items referring specifically to the degree to which an attitude is accepted or rejected. An individual may check items having his approval, whereby he may state, for example, that he is "heartily in favor of the proposition," "has no special objection to it," "does not believe it affects him greatly," or "is strongly against it."

There are other methods by which an individual's attitude may be evaluated. Most of these represent attempts to disguise the attitude-object by not mentioning the attitude or any stereotype associated with it. It is believed that antagonisms or prejudices are aroused by mention of such terms as "socialism," "fascism," "Reds," and "prohibitionists."

Sociometric diagrams have been used as a basis for making inferences regarding social prejudices. Within a group, each member may be asked to designate those whom he would accept as a citizen, as a neighbor, as a member of his family. The assumption is that each individual is guided by certain social prejudices for or against others. The "cross-out method" is a form of projective technique, by which the individual is given a series of terms and instructed to delete those that are unpleasant to him. It is a means of determining whether he has any consistent prejudices.

Evaluation of attitudes tends to be made on the basis of behavior which often is not highly specific to each attitude. Indirect methods tend to be more accurate than direct methods in determining the direction and intensity of an attitude. If the individual is asked to display behavior highly specific to an attitude, he may simulate unnatural behavior in order to reveal what he believes to be the desired response. In order to determine true attitudes, it is frequently necessary to use a technique which will disguise the attitude-object so that an individual's spontaneous reactions will be elicited.

SOCIAL COMPETENCY

Social competency is implicit in all statements of broad educational outcomes. It is selected for discussion because it throws light upon the complexity of many new and challenging frontiers of evaluation in education.

Social competency refers to the many abilities which relate to effective human relationships. Even the value of educational outcomes which appear restricted to an individual's self-realization—

such as language and mathematical skills—prove upon analysis to depend upon the social utility of such skills. Ability to write, read, listen, and express thoughts orally possesses value only within an appropriate social environment. Ability to use the English language, for example, is of value only in an environment in which persons comprehend it.

Social and vocational needs. All school subjects presumably contribute to development of ability to satisfy human needs. The social studies share this common purpose and also deal with a content specifically related to problems of social living. Much research in the field of the social studies has been conducted in order to discover the specific social needs of individuals, the activities which are required for social competency, and their classification and generalization as objectives of instruction. An early study by Wilson[1] consisted in having 4,068 adults keep careful records of the types of arithmetic problems with which they had to deal during a two-week period. Analysis of such problems yielded information as to kinds of arithmetical processes needed in daily life and their degrees of complexity. Other areas explored by research have included necessary geographical names, the practical use of chemistry, and types of undesirable behavior as revealed in police court records.

Many recent studies have dealt with job analysis. Their purpose has been to discover the skills and knowledge necessary to engage in various types of vocational activity. It has been believed that information thus derived might afford a basis for providing appropriate instruction and also be valuable in counseling. Studies of trends in science, home, social life, and in governmental activities have also afforded a foundation for revision and further exploration of the types of training regarded as socially important. Children in the schools receive information concerning the values of different foods as a result of advances in nutritional science. The "pioneer-writer technique" has been widely used. This technique consists in observing the degree of importance accorded topics in various fields by authoritative writers.

The problem of determining social values has sometimes been approached negatively. Investigators have sought to identify types of social incompetence and to determine causes which might relate to backgrounds of training. One investigator analyzed the behavior of individuals who were regarded as failures as citizens, husbands,

[1] G. M. Wilson, *What Arithmetic Shall We Teach?* (Houghton Mifflin Company, 1926).

or as cultured individuals. Such investigations afford insight into the nature of educative experiences designed to prevent social maladjustments.

Determination of needs and of social values might result in nothing more than increase in knowledge. However, knowledge of what individuals need to know how to do often may stimulate reappraisal of educational outcomes and the means of their attainment. Knowledge also makes possible more realistic interpretations of the objectives to be sought in terms of specific behavior.

Evaluation of the attainment of accepted educational outcomes may also involve evaluation of the outcomes themselves, inasmuch as the desirability of many kinds of social behavior is a closely related problem. While the broad outcomes are being analyzed to yield various types of specific behavior, the question continually arises: Is the type of behavior importantly related to the individual's social competency?

Social needs may disappear unaccompanied by changes in forms of training, just as they may appear before forms of training have been adequately developed for satisfying needs. Almost forty years ago an observer cited evidence that ability to perform the arithmetical process of extracting cube root was of no practical value for most individuals, although the process was still being widely taught during the waning days of the mental-discipline philosophy. Utility of abstract mathematics for many individuals is currently debated, as is also the value of foreign languages for the large number of individuals who rarely would have occasion to use any foreign language studied. Thus the evaluation of outcomes may reveal learning of little social significance and call attention to neglected knowledge which would contribute substantially to social competency.

Curriculum planning. Educational outcomes are of direct concern in the development of the curriculum. The curriculum planner is concerned primarily with selection of appropriate subject matter. His purposes cannot be successfully accomplished unless he takes into account the kinds of learning desirable for certain groups of individuals. He must also consider whether certain kinds of learning are likely to result from study of a certain type of subject matter. He must appraise the needs of individuals at different ages, the needs for a particular community, and the needs of those vocations which individuals in his locality most frequently enter. To the extent to which such ultimate outcomes are considered when instructional objectives are formulated each teacher becomes a curriculum-maker.

Because of the dominant position of social competency, the question may be asked for each course of instruction: What information, skills, and attitudes may result from instruction that will be of the greatest social value for the greatest number of individuals? What should individuals become able to do that they were unable to do before instruction? Instructional objectives may be developed so that they consistently support the broad social goals of education.

Behavior observed with reference to social competency is largely that associated with recall of information or use of information. We are not likely to achieve social competency as a result of attempting to help a child in his struggle with the difficulty of a certain aspect of addition, especially if the teacher's attention is concentrated almost exclusively upon the immediate problem. The contribution at the time, however, is not without value. It serves to create understandings that, as they accumulate, may aid the child in satisfying certain social needs in the future.

The preponderance of the child's practice in using addition to satisfy his social needs occurs outside the school, that is, in his home or at the grocery store. Practice in the classroom tends to be concentrated upon minute segments of learning. Pupils in the schools achieve effectiveness in social living when teachers constructively guide the relationships of pupils with each other and with their teachers. Good results have been obtained when such relationships were used so that the school became in many respects a replica of a small community.

A study of the scales developed by Peters[1] reveals many components of social competency. Through use of such scales, the extent and form of an individual's social competency may be estimated. An excerpt from these scales is presented:

INVENTORYING SOCIAL COMPETENCE

Nearly perfect	Greatly	Considerably	Moderately	Slightly	None[2]
●	●	●	●	●	●

I. A Blueprint of Personal Culture

A. The cultured individual has ability to get joy out of living

1. Ability to enjoy various forms of art and nature

a. ability to enjoy music

(1) ability to enjoy some music or other

[1] C. C. Peters., *The Curriculum of Democratic Education* (McGraw-Hill 1942. p. 297 ff). Reprinted by permission.

[2] The large dots are repeated at intervals throughout as a device by which individuals may be rated.

II. A Blueprint of an Optimum citizen

A. The efficient citizen should be prepared to maintain proper perspective as to his place in organized society
   1. He should have become so habituated to the social outlook that he measures all conduct from the point of view of the impartial spectator
      a. He should be disposed and able to hold his rights as an individual no higher relatively than those of others

B. The efficient citizen should be able to perform effectively his political obligations
   1. Participate effectively as a lawmaker
      a. He should have a sense of his personal responsibility for government, a recognition of his obligation to participate in all of its functions in which he is given a voice, and interest in civic affairs.
      f. He should have command of the necessary tools of self-help for getting information needed in his civic activities.

III. A Blueprint of Vital (Physical) Efficiency

IV. A Blueprint of the Domestically Efficient Person

V. A Blueprint of Vocational Efficiency

VI. A Blueprint of Social Democracy

VII. A Blueprint of Industrial Democracy

From so comprehensive a list of statements outlining desirable generalized behavior, upon what basis are certain objectives to be selected in order that specific acts may be defined and taught? In the light of present knowledge and experience, only an incomplete answer is possible. It should be pointed out, however, that the role of instructional objectives varies with the plan of instruction. When conventional methods are used instructional objectives represent types of behavior that practice on learning material is expected to create or improve. Such objectives represent preconceived notions of aims to be reached during a course.

There is a tendency of educators to profess desire for certain goals and yet fail to provide educational experiences that will lead toward their attainment. A course in physical science may be expected to "develop familiarity with scientific method." Yet many pupils, upon completion of the course, do not have an adequate conception of the nature of scientific method. The course may have resulted in no gain in ability to approach problems scientifically. When conventional methods are used there is always danger that

attempts to obtain improvement in certain aspects of social competency may result only in academic learning.

Conventional methods, for example, make provision for only a few kinds of behavior related to the practice of social competency. The generally conceived purpose of problems referring to use of arithmetic in life situations is to clarify arithmetic and make it "meaningful." Aspects of social living introduced to clarify subject matter thus serve primarily as a means to an end. Difficulty in focusing instruction upon problems of social competency is inherent in conventional methods of teaching. If any goals that are closely related to social competency are reached such results are generally due to incidental learning.

Considerable progress has been made, however, toward achievement of broad educational outcomes by making instructional aims more functional. During the past two decades a marked refocusing of emphasis has been evident in the teaching of history and civics. Instructional aims for history have been broadened so that they take into account the attitudes formed by pupils, present the field as a means of understanding the present by understanding the past, and emphasize history as a record of social change. Civics no longer describes impersonally the structure of government but attempts to clarify the functions of government and their relationship to citizenship. An important aim is to prepare pupils for functional roles in society. Although use of language and similar sources of vicarious experiences are the predominant means of demonstration, certain activities are carefully planned. These activities are centered upon actual pupil participation in miniature civic situations. As a result of such experiences pupils gain sensitivity to some of the realities of citizenship. The objectives of social competency are found to be less visionary than supposed. Promising results have been obtained.

ACTIVITY PROGRAMS

Under an "activity" plan of instruction, method is not restricted by subject-matter fields. Instead, various programs of the school are centered upon the relatively broad outcomes of education, with emphasis upon needs appropriate for various age groups. Groups of pupils representing various ages and degrees of maturity may develop their own units of study. Each unit may represent some important area of social living. The activities of such groups are frequently regulated by socialized procedures. Discussion may

begin with suggestions by pupils of problems related to a central theme. The unit might be *How We Obtain Our Food.* Eventually, there may be developed a long list of subtopics which relate with increasing specificity to the central theme. Discussion of different kinds of food, their sources, distribution, prices, and similar topics readily serves to emphasize the need for additional information. Need for examining the contributions that may be obtained from various sources in other subjects becomes apparent to the pupil. Not only are the general themes related to problems of social living, but co-operative effort in obtaining and organizing information is a desirable type of practice in social living.

In "activity" programs the objectives emerge and become meaningful to the learner while a course is in progress rather than being specifically formulated in advance as goals eventually to be reached. Information in the subject is introduced when it can contribute to the attainment of these emerging objectives. Attention to information is not permitted to dominate the social purposes of the "activity" program. Social outcomes are realized when members of a group work together in analyzing and interpreting relevant contributions of formal subject matter.

A democratic, activity-centered plan developed by Peters[1] and experimentally tested in eight high schools illustrates possibilities of practicing behavior that characterizes good citizenship within the social studies as a frame of reference. It was explained to a co-operating group of teachers in civics how instruction which affords practice in life behavior differs fundamentally from that which is involved in mastery of school subject matter as such. The teachers were informed that the instruction would stress personal choices, habits, and programs of personal actions.[2] A selected list of 25 articles (reprinted from *Reader's Digest* and dating from 1927) relating to citizenship in action was used as introductory discussional material. The early part of the course was devoted to accustoming the group to informal discussion and to socialized procedures similar to those of a conference, with all persons being co-learners and co-workers.

Constructive progress was then begun toward developing "citizenship" as a central theme, first by thinking of some of the activities that citizens ought to be able to perform. Suggestions were written on the blackboard. The full range of citizenship was pointed out as including not only "political citizenship" but also

[1] C. C. Peters, *Teaching High School History and Social Studies for Citizenship Training* (Coral Gables, Florida, University of Miami, 1948).
[2] C. C. Peters, *op. cit.*, p. 52.

school, home, and community living as these go on now and will go on in the future."[1]

Several days were required for analysis of behavior. Following the initial attempts to analyze citizenship, copies of *Peter's Blueprint of an Optimum Citizen* were made available as a basis for checking and extending the analysis. After several weeks it was anticipated that approximately 340 specific elements relating to citizenship behavior would be formulated. Instructions to the group for dealing with such elements were as follows:

> . . . Take up the first item in class: 1-a, "He should be disposed and able to hold his own rights as an individual no higher relatively than those of others." Notice that this is the Golden Rule. Do Americans now behave that way? Would it be practical? What are some of the things people do in violation of that principle? Some things you have seen people do in conformity with it? How could one train one's self into actually carrying out this philosophy in his own life?[2]

As the discussion continued throughout the course, each pupil rated himself as to how fully he believed he was developing in each trait. In no case was a self-rating final. On the basis of a fuller understanding of the meaning and implications of the items, pupils were encouraged to revise their ratings continually. Upon completion of the course each pupil had a profile of his individual characteristics as a citizen showing his points of strength and weakness.

Two general goals were emphasized simultaneously in the experimental groups, whereas in the control groups only the conventional goals of the social studies were stressed. The major training toward "citizenship" in the experimental groups consisted not only of analysis and study of citizenship but of the group practices followed throughout the discussion. Various specially designed tests were administered to both groups as well as standardized tests. In some cases instruments of the "Guess Who" type were used. This type of exercise made it possible to obtain ratings by the pupils serving in the capacity of observers of their own and each other's behavior. A portion of a test used in a similar democratic, activity-centered experiment suggests the detail with which each individual's behavior was examined by his peers. The test contained 29 propositions, of which the following are samples:

> There are some members of this class who feel responsible for the success of the class, in such matters as good discussion and reports, good

[1] C. C. Peters, *op. cit.*, p. 54.
[2] *Ibid.*

reputation of the class before the school and community, etc. *Who are they?*

Some of the members can express themselves convincingly, so that other pupils are won by their statements or arguments. *Who are they?*

Other members carry little conviction when they speak; do not convince or persuade others. *Who are they?*

Some of the pupils are tolerant of the opinions of others; they criticize other people's views with respect. *Who are they?* [1]

These propositions were classified under five broader behavior patterns:

1. Feel responsible for the success of the group.
2. Manifest leadership.
3. Have an interest in learning and make progress in learning.
4. Are co-operative, orderly.
5. Are tolerant, courteous, considerate of others.

The number of nominations each pupil received as an aggregate under each of these types was taken as the index (score) of his behavior in that trait.

Peters believes that pupils taught by the methods just described are not inferior to those in the conventional academic masteries and are superior in initiative and in ability to solve life problems in their areas of training. In general, achievement in citizenship on the part of the experimental group became definitely superior with successive semesters of instruction. This finding suggests a progressive increase in the understanding of behavior related to citizenship and in the experience of the teachers who used these novel methods.

It should be emphasized that only part of the evaluation of any type of achievement has been completed at the conclusion of a period of instruction. Evidence obtained during training periods of behavior identified with good citizenship can be only presumptive evidence that the individual will continue to manifest such behavior during the remainder of his lifetime. In order to determine reliably whether habits are firmly established it would be necessary for an observer to study the individual during several periods of his lifetime and to record his behavior in all of its significant forms.

The most important conclusion for evaluation is the fact that even with imperfectly designed instruments we can detect evi-

[1] C. C. Peters, "An experiment with democratized education," *Journal of Educational Research*, 1943, *37:* 95–99.

dences of progress toward almost any objective we wish to formulate. A statement by Peters is significant:

> A vast amount of experimental evidence has been accumulated during the past 40 years that shows we can achieve objectives in education by specifically working for them . . . leadership and courtesy and international-mindedness, and other traits can be measurably increased by making them the objectives of well-planned educational efforts. In the experiment . . . we set as one of our objectives enough of the conventional academic masteries that our experimental pupils would not fall too far behind the control pupils—and we achieved that objective. We could doubtless have achieved more or less of this according to the price we were willing to pay for it. We worked for civic objectives; we could, instead, have worked for cultural or health or other objectives. We can achieve with our pupils docility or independence, thinking or rote memorizing, broadmindedness or provincialism, functioning social abilities or drawing-room erudition, according as we make one or the other of these pairs the conscious objective of our educational set-up and drive purposely toward it. Without *any* conscious specific objectives (and much teaching has no objectives other than to get through the year and through "the book") little if anything is likely to be accomplished.[1]

SUMMARY

The broad comprehensive terms which are used to describe general human traits and educational outcomes tend to create confusion and disagreement as to meaning. The first problem in evaluation is to delimit the area which is to be evaluated, so that the delimited area will possess unity and homogeneity. The second problem is to decide upon behavior both generalized and specific which may serve as criteria for determining whether progress toward definite aims has occurred. Evidence for possession of a trait or attainment of an instructional objective is inferred from the individual's behavior. The general method of evaluation is to formulate instances of characteristic behavior and to evaluate the extent to which such behavior is displayed.

Presumably all outcomes in specific areas should contribute consistently to the general goals of education discussed earlier in this chapter. All possibilities for influencing individuals should be examined in connection with a subject-matter field. Every educational activity, even if relatively independent of a subject matter field, should be similarly scrutinized.

[1] C. C. Peters, *Teaching High School History and Social Studies for Citizenship Training* (Coral Gables, Fla., University of Miami, 1948, p. 141).

The research worker's initial activity in defining educational outcomes is to create numerous behavior situations, representing each broad category of objectives tentatively chosen for study. If the investigator takes into account the possibilities of given situations he not only defines his objectives clearly in terms of behavior specific to the objectives but also provides the groundwork necessary for an evaluation program.

Outcomes are satisfactorily defined only when analyzed in terms of the behavior of which the individual becomes capable as evidence that the objective has been achieved. Formulating a number of specific behavior situations creates a valid foundation for evaluation. Numerous specific restatements of objectives in the form of specific acts serves as a safeguard against formulating objectives that are incapable of evaluation.

CHAPTER III

Quantification of Educational Data

Quantification refers primarily to any determination of value expressed in numerical form. Quantification of different values relating to the qualities of some person or object leads to consideration of a restricted number of such qualities. This is the basis for the frequent observation that quantitative methods do not reveal a comprehensive picture of many important qualitative features.

We may wish to quantify a child's achievement in arithmetic. In order to do so, we test his ability to perform numerous specific acts requiring knowledge of arithmetic. Although we tentatively regard the total number of correct responses as an index of achievement, the child's knowledge may have developed in directions not taken into account by the test administered. Thus, our success in quantifying his achievement is limited by the extent to which we have been able to translate our concept of achievement into specific aspects of behavior.

As another example, we may compare two textbooks which are unlike in almost all respects. Many qualities of each book such as length, width, thickness, weight, number of pages, number of chapters, or number of words may be quantitatively expressed. Yet, each quantitative operation is essentially abstractive. Use of measuring techniques not only results in loss of the individual features of each quality but affords only abstract quantitative definitions.

Verbal statements asserting existence or nonexistence of a given quality are often sufficient for satisfying demands of accuracy in daily conversation. But when discussion turns to differences in the

extent to which a quality is shared among individuals, qualitative statements concerning observed differences are inadequate.

Since quantitative methods make it possible to derive many important numerical representations of qualities, our initial discussion is concerned with the role of number. In the second part of the chapter we shall point out how numerically expressed values emerge in connection with various types of data-gathering devices. Discussion of refined techniques for dealing with group data and of qualities which data-gathering devices should possess is reserved for the next chapter.

THE ROLE OF NUMBER

A number has limited meaning in the absence of appropriate context. Unwarranted inferences may be avoided by studying carefully the situation in which a number is used. We may cite several instances in which confusion may occur.

Classrooms in a school are ordinarily assigned numbers. A room numbered 200 enables us to identify it. Inasmuch as the number 200 serves only to lend distinguishability, it may not be inferred that (*a*) the school has at least 199 other rooms in addition to room 200; (*b*) room 200 is twice as desirable as room 100; (*c*) room 200 is larger than 100; or (*d*) all rooms are used for similar purposes.

The term "three inches" may be used to express the length of an eraser or of a fountain pen. No concrete object constitutes an inch, and neither eraser nor fountain pen is three inches. In order to make a strictly precise statement, we would have to say that the eraser or the fountain pen is "as long as" or possesses the "same length as" three inches. These objects possess length but do not possess inches.

Some of the methods of applying number for the purpose of expressing quantity will be considered.

Counting. The simplest method of quantification is that of counting. One's interest may be not only to verify the existence of a certain opinion among individuals in a given population but to quantify the extent to which the opinion is held. The quantification operation will be that of counting the individuals sharing the opinion and the individuals sharing an opposed viewpoint. The technique of using the method may consist of a personal approach to members of a population and determination of their opinions by direct questioning. Information may also be obtained by use of a written questionnaire. All that is involved, regardless

of specific technique, is a counting of people. We do not actually quantify opinion or seek to express it in specific amounts but use the number of persons holding an opinion as a quantification of the extent to which it is held.

In counting, two basic properties of a number are involved. A number may lend distinguishability to a person or object or may serve to quantify a plural concept. In counting the number of persons in a room, we assign to each person a number in a standard series. The typical standard series is that of consecutive whole numbers arranged in ascending order. The first person might be designated number 1, the second as number 2, and so on. Each number thus upon assignment may serve to distinguish an individual from the other members of the group. The quantity of persons, or the index of plurality with respect to the group, can be determined by using in another sense the largest serial number assigned. Number 85 may be an attribute of the last person (the numerical index of quantity is 85), each member of the group having served as a unit of quantity. Number 85 may represent the entire group, and the counting process has been completed, since the largest possible number has been reached.

Counting is possible only when one is dealing with an aggregate of *discrete* units. We may count the number of high schools in a state which offer three years of French, or the books in libraries. In such situations, each item is distinct.

Footrule method. In the preceding discussion, quantification was achieved by the counting of persons or objects which remained essentially invariant. These were dealt with as discrete members of a standard numerical series. It is frequently necessary, however, to quantify certain qualities which are continuous in nature.

For one type of continuous quality, a footrule method may be applied. By means of this method, use is made of standard scales. Such scales may be applied to certain qualities of persons or objects. The standard units of physical measurement include those of length, volume, degrees of angles, etc.

Upon a footrule, not only are distances from zero so marked as to indicate ascending order of magnitude, but each numbered division is numerically proportional to the actual distance from the zero end of the scale. Thus, number 6 (interpretable on a footrule as 6 inches) expresses the fact that the division so marked is *sixth in order* of the one-inch divisions and may also be interpreted as the total distance from the zero end.

For practical purposes, we quantify the object just as though the

object itself were divided into inch units and could be measured by inspection of its extensive qualities—length. Footrule measurement may be regarded as direct measurement by means of some physical instrument.

Rank order. For purposes of making comparison among persons and objects with respect to qualities, the rank order method is often convenient. Sometimes the rank order method is the only method by which valid comparisons may be made. In some instances, it constitutes a final court of appeal in interpreting the quantified results of the counting or footrule methods.

"Smoothness," "brilliancy," "diligence," and "enthusiasm" are qualities which have been referred to as existing in a continuum of amount. We can not make a direct quantitative estimate of the enthusiasm of one individual in comparison with that of another.

Rank order is highly important in the quantification of educational data. It is the only possible method of quantification of many traits or qualities possessing educational significance. With appropriate refinement it is a method of translating raw test scores into quantified forms in which they may possess meaning and usefulness for the research worker in education. The basic techniques for interpreting the results of rank order measurement will be considered later in connection with typical educational applications.

Derived measurement. *Derived* measurement is similar in many respects to indirect measurement in the sense that standard units are not applied directly to what is measured. Most typically, derived measurement occurs whenever two or more cases of fundamental measurement are combined in ratios which describe a relationship. The effect of the relationship may often be felt, but its measurement can not be made without a calculation which is based upon plurality of values. School marks, for example, not only often take into account various types of performance which a learner can exhibit but also either the relationship of such achievement to expected achievement or the rank value of each individual's achievement among that of other individuals. Many different circumstances must usually be considered in assigning school marks. Derived measurement is, in general, a result of resort to relatively remote correlates and usually to a multiplicity of correlates.

In the field of education, few natural laws can be conveniently utilized for making a derived measurement. There is, however, some dependence upon the mathematics of logical relationships. The number representing the IQ is "derived" from two measure-

ments or quantified observations. One is the individual's mental age as determined by means of standardized behavior typical of "normal" performance at different ages. The other measurement consists of the calculation of chronological age in terms of months. The IQ is the quotient obtained by dividing the mental age in months by the chronological age. If both values used are numerically indentical, the IQ is 1.00 or (100 as written for convenience.) We may speak of the IQ as a derived measure since it refers to an abstraction based upon an empirical relationship.

In the case of derived measurement in the field of education, the exact functional relationship between the values used in order to quantify such a relationship is generally uncertain. Although the individual's chronological age in months is involved in the determination of his IQ, chronological age in itself is significant only because in studies of intelligence the aim has been to determine normality of intelligence for individuals of various chronological ages. Of all aspects of change occurring during an individual's life, the rate of change in chronological age is obviously invariable. There is no causal relationship between chronological age, however, and the height of an individual's mental growth. It is generally agreed that the quality of mental growth is not invariable during a lifetime—that patterns of thinking, for example, developed during adulthood differ from the thought patterns of youth. The qualities and traits measured in education and psychology by derived measurement do not possess additive properties as is the case in many physical qualities quantified by derived measurement. We may add kilowatt-hours, for example, but it is not possible to calculate the mass value of the measured intelligence of a group of individuals.

DATA-GATHERING DEVICES AS INSTRUMENTS OF QUANTIFICATION

For the research worker any activity manifested in devising or selecting a technique for gathering data is identified with a purpose. Such a purpose may be expressed in a carefully formulated problem, or its existence may be implied by the fact that the research worker selects a general area in which to initiate data-gathering activity. It is conceivable that data gathered may result in bringing a problem which is only vaguely felt into sharp focus and suggesting the need for revision of procedures and often collection of new data. But, in general, a data-gathering device is selected or designed with reference to a problem which is felt.

Within the framework of the research method, the research worker must select or devise a technique of gathering data which is appropriate to the sources from which data may be gathered. Use of a questionnaire may be appropriate to some sources of information, whereas administration of tests may be appropriate to others. The technique of collecting data refers primarily to the manner of employing the method. The method must be used in such a way that the data obtained are adequate in kind and form in order to be analyzed with reference to the problem. The research worker who believes that relevant data may best be obtained by means of a test may discover that he must design one which will adequately quantify whatever he wishes to measure. An available test of range of information, for example, may not be satisfactory for measuring ability to interpret information.

The purpose of our present discussion is to clarify the nature of the process by which various typical data-gathering devices provide a means of deriving numerical indices of value and to point out how such indices are correlated with the educational values studied. Many techniques of obtaining data are not inherently capable of yielding quantified information initially useful for research. In fact, verbal communication and display of behavior are not intrinsically self-quantifying, although they serve to sustain quantification. No idea of number, for example, resides in the information presented when an individual answers two questions on a history test. The qualitative aspect manifested in the correctness of the information however is reflected in the fact that there are *two* instances of correct performance. Here number is not a direct quantification of quality but essentially an index of frequency. In considering typical data-gathering devices, emphasis is given to many controls which are usually necessary in order that numerical values may not only emerge but be comparable in variability with that possessed by the aspect of the quality studied.

It is not possible in limited space to give encyclopedic treatment to the almost infinitely varied types of quantification techniques which have been used to collect and develop data in education. For purposes of discussion and illustration several general headings are considered: (*a*) observation, (*b*) the critical behavior technique, (*c*) interview, (*d*) questionnaire, (*e*) inventories, (*f*) rating techniques, (*g*) rank-order scaling, and (*h*) testing situations.

Observation. Broadly interpreted observation refers to any act of obtaining information, such as that of measuring a table with a

scale or of interpreting a pencil-and-paper test. In a restricted sense, as used here, observation implies that an examiner's various senses are functionally involved in the acquirement of information. Restriction of the term "observation" is made for convenience in discussion, even though it may still refer to various types of aided and unaided perception. It may be said that an observer is in close proximity to his sources of data or to behavior witnessed and that he records what he perceives without extensive use of measuring instruments. In certain instances he may be aided by a precision device, such as a stopwatch, or he may use some mechanical means of counting examples of behavior. But in such cases he uses aids to observation and is personally responsible for his results, in much the same way as a medical student who uses a microscope. An example of observation may be the activity of an individual in watching children at play and familiarizing himself with the types of play in which they engage.

If personal observation is not fortified by some device for controlling the influence of time or by some check-list specifying what the observer is to make note of, observation may result in diffuse descriptive data which may include the effect of personal interpretations or emotional reaction. For research purposes, the "human factor" must be minimized in the interest of objectivity.

Exploratory observation, however, often may precede establishment of controls defining in detail how subsequent observation will be made, what specific items the observer looks for, and how data obtained will be reported. Suppose one were to observe and report on the reading habits of eight-year-old children. If there were no systematic guides to follow in making observations of such behavior, or if there were no established criteria for characterizing observed behavior as "good" or "bad," exploratory observation would be necessary in order to determine the reading habits of such children. The observer might decide to obtain data concerning the frequency with which the children (1) mispronounce words during oral reading, (2) omit words or entire lines in oral reading, (3) vocalize during silent reading, or (4) point to each word while reading it silently. A detailed outline of behavior to be observed is indispensable if instances of performance are to be quantified. Sometimes exploratory observation results in detailed verbal descriptions which may be used as a basis for constructing a check list.

Certain conditions which affect the accuracy of an observer's activity are outlined:

1) *The observer should possess efficient sense organs.* An individual who is hard of hearing would be unable to detect all instances in which children mispronounce. A nearsighted person would be handicapped in the study of children at play.

2) *The observer must be able to estimate rapidly and accurately.* He may be required to decide whether a certain child is extremely shy toward other children on the playground or whether only mildly so. He often must estimate quantitatively assigning rating values, for example, to certain characteristic behavior of the child.

3) *The observer must possess sufficient alertness to observe several details simultaneously.* He must be capable of sustained attention in order to remain aware that certain behavior is occurring continuously. Individuals, as a rule, are very sensitive to fluctuations in behavior. We detect the presence of moving objects more readily than of objects in a state of rest.

4) *The observer must be able to control the effects of his personal prejudices.* He must record observation on the basis of what he actually sees or hears and not on the basis of what he believes he should see or hear.

5) *The observer should be in good physical condition.* His accuracy in observing may be detrimentally affected by loss of sleep or fatigue, thus rendering his attention ineffective for making sustained observation. A nervous or overwrought individual may base his report upon illusory sensory responses.

6) *The observer must be able to record immediately and accurately the results of his observations.* Postponement of recording an event for a few minutes may increase the likelihood of inaccuracy.

It is often desirable to determine whether results of observation are free from personal bias or human error. One method, which consists in requiring observers to repeat their observations, affords opportunities to reveal certain errors which have resulted from personal inconsistency.

A method involving the services of several independent observers is usually regarded as the best safeguard against personal bias. Even a professionally trained observer may impose his personal points of view upon phenomena observed. As a rule, if the results obtained by several observers, each working independently, are pooled and if no substantial disagreements are found, the results of observation may be regarded as reasonably reliable for many purposes. If a single member of a group of observers reports a find-

ing conspicuously inconsistent with that of the majority of observers, the probability of bias or error may be suspected. If all observations disagree substantially with one another, it is improbable that valid conclusions may be made from the information available. In any case, it is important to estimate the probability of chance variation in the consistency of the performance or behavior observed. These observations are applicable generally in any gathering of data and are especially significant in the case of rating methods.

An example of observational technique involving time sampling appears in a study by Gesell and Ames [1] concerning the incidence of definite "handedness" in the case of young children. Their investigation was based upon periodic observations of a small group of infants and children as their ages increased up to 10 years or more, with emphasis upon the earliest years of life. Comparable consecutive data were obtained from cinema records at lunar month intervals for the first 60 weeks of life and at less frequent intervals thereafter. These data were combined with stenographic observational records, in order to make possible a detailed tabulation of individual performance by seven children whose records were complete through the age of ten years.

Table 1 presents in schematic chronology the characteristic age shifts in the handedness of subjects, all of whom eventually showed definite, clear-cut right-handedness. Figure 1 illustrates these shifts as they were observed in the seven basic cases for whom cinema records were available through the first 10 years, and one additional case through the first two years. The figure shows at each age level the number of cases in whom right-handed, left-handed, or bilateral behavior occurred predominantly or very conspicuously at each monthly level up to 60 weeks of age, and at yearly age levels from 2 to 10 years. The totals indicated on the graphs are frequently more than eight, because some cases definitely exhibited more than one type of handedness at a given age.

It was concluded that the children made contact with objects first with the nondominant hand, then bilaterally, then with the dominant hand alone, once again bilaterally, and then with one hand—usually, and to an increasing extent, the hand which was dominant. There appears to occur in children at 1½ years of age a period of marked bilaterality, followed by consistent use of a dominant hand alone at the age of 2 years. Another period of bilaterality occurs between 2½ and 3½ years, but from 4 years on-

[1] A. Gesell and L. B. Ames, "The development of handedness," *Journal of Genet. Psych.*, 1947, *70:* 155–175

TABLE 1. Schematic Sequence of Major Forms of Handedness

16–20 weeks:	Contact unilateral and, in general, tends to be *with left hand.*
24 weeks:	A definite shift to *bilaterality.*
28 weeks:	Shift to unilateral and oftenest *right hand is used.*
32 weeks:	Shift again to bilateral.
36 weeks:	Bilaterality dropping out and unilaterality coming in. Behavior usually characterized 'right or left.' *Left predominates in the majority.*
40–44 weeks:	Same type of behavior, unilateral, 'right or left,' but now *right predominates in majority.*
48 weeks:	In some a *temporary, and in many a last shift,* to use *of left hand*—as well as use of right—either used unilaterally.
52–56 weeks:	Shift to clear unilateral dominance of *right hand.*
80 weeks:	Shift from rather clear-cut unilateral behavior to *marked, interchangeable confusion. Much bilateral and use of nondominant hand.*
2 years:	Relatively clear-cut unilateral use of *right hand.*
2½–3½ years:	Marked shift to *bilaterality.*
4–6 years:	Unilateral, *right-handed* behavior predominates.
7 years:	Last period when *left hand, or even both hands bilaterally,* are used.
8 years ff.:	*Unilateral right* once more.

ward the dominant hand is used increasingly. In some cases, even at the age of 7 years, there is temporary use of the nondominant hand or of both hands together.

The basis for quantification depended on the fact that instances of behavior could be quantified by counting and that evidence for any pattern of handedness might be found in frequency of occurrence. No attempt was made to demonstrate relative value of performance among the children studied.

Critical behavior technique. The critical behavior technique [1] is a relatively new development in research. Among the more

[1] This technique, developed principally by John C. Flanagan, has been used effectively both in military and civilian situations. For a clear statement of the purposes and procedures involved see *American Institute for Research, A Report of Three Years of Experience* (American Institute for Research, 413 Morewood Avenue, Pittsburgh, Pennsylvania, September 1950).

See also John C. Flanagan, "Research technique for developing educational objectives," *Educational Record,* 1947, *28,* 139–148; and John C. Flanagan, "Critical requirements: a new approach to employee evaluation," *Personnel Psychology,* 1949, 2 No. 4, 419–425.

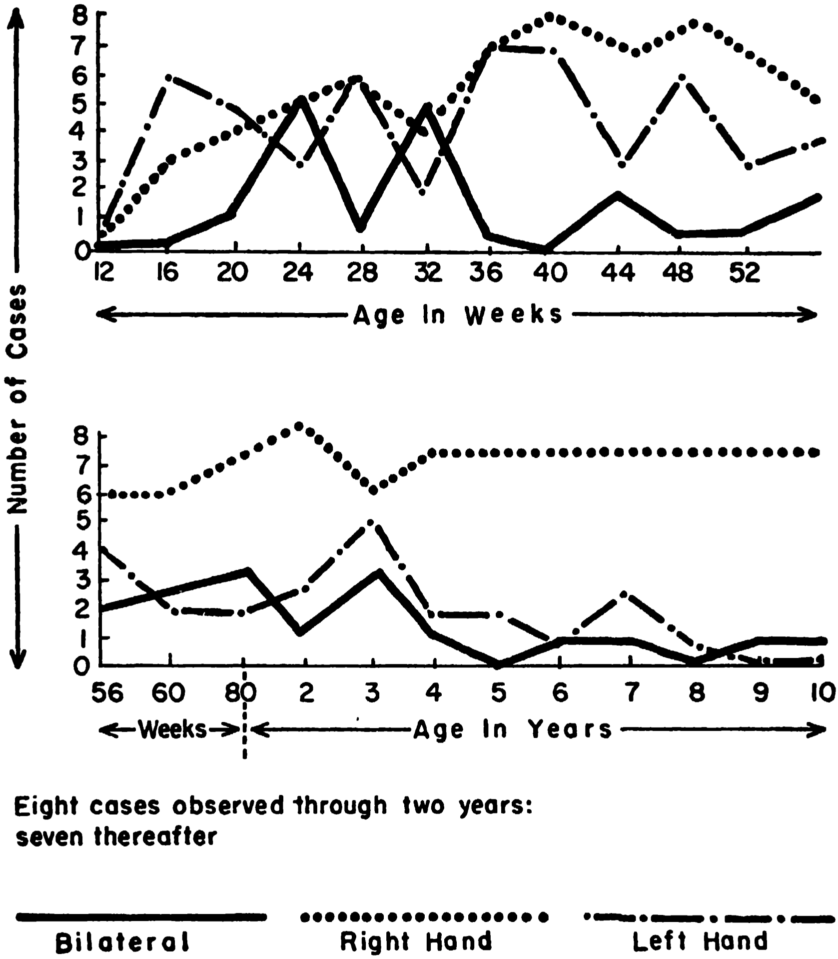


FIGURE 1. A study of handedness. (From Gesell, *Journal of Genetic Psychology*, 1947, 70:158.)

important features of this method is the systematic collection and analysis of factual data, rather than reliance upon opinions, impressions, and estimates. It involves the formulation of a comprehensive list of traits that have been observed to make the difference between success and failure in a given activity. It appears to be a very promising approach to the establishment of criterion measures and the formulation of educational objectives.

If valid results are to be obtained certain specific conditions must be satisfied as follows:

a) It is essential that actual observations be made of the on-the-job activity and the product of such activity.

b) The aims and objectives of the activity must be known to the observer. Unless this condition is fulfilled it will be impossible for the observer or judge to identify success or failure. For example, a foreman might be rated as very successful if the objective of his activity were taken as getting along well with the workmen under him. At the same time, he might be rated as very unsatisfactory if the objective were to produce materials.

c) The basis for the specific judgments to be made by the observer must be clearly defined. The data can be objective only if all observers are following the same rules. All observers must have the same criteria for making judgments. For example, the definition must clearly state whether or not a minor imperfection will be regarded as evidence of failure or whether a product must be completely unusable to be classified as unsatisfactory.

d) The observer must be qualified to make judgments regarding the activity observed. Typically the supervisor on the job is in a much better position to make judgments as to whether behavior is outstanding or unsatisfactory than is the job analyst or psychologist. On the other hand, the supervisor on the job is ordinarily lacking in the training essential to make an inference as to the particular mental trait which caused the behavior to be successful or unsuccessful.

e) The last necessary condition is that reporting be accurate. The principal problems here are those of memory and communication. It is also important that the observer's attention be directed to the essential aspects of the behavior being observed.

The interview. Information which the interviewer obtains tends to be of a qualitative nature. The subjective nature of the interview may be an asset when used in counseling, personnel placement, or clinical psychotherapy. For many clinical purposes the uncontrolled interview is more desirable than one which restricts the individual's freedom of expression. In the uncontrolled interview the direction of questioning is continuously governed

by (1) the examinee's immediate reactions, and (2) the nature of the information elicited during the interview. Under such conditions many advantages accrue from interpersonal relationships that would not be possible when using other techniques.

For some problems, the interview technique may result in uneconomical use of time and energy, since the emphasis is predominantly upon a person-to-person relationship. If the interview is used in obtaining information concerning a group, each member must be approached individually. It is difficult to quantify information when the interview is used unless the questioning procedures are carefully designed and rigorously followed. Yet the interview should not become formalized to the extent of reducing the values which may result from rapport between the interviewer and those whom he interviews. The interview should be restricted to the types of information desired and at the same time be flexible in permitting certain individuality of expression.

For developing quantitative data it is important to outline in advance the kinds of information desired. The principal function of the interview then becomes that of obtaining responses which may be recorded under various categories. In many respects, the interview so planned is essentially an oral questionnaire. Yet its oral aspects make its use desirable in the case of (1) young children who would be unable to write replies to a questionnaire or upon a written test, and (2) types of information the securing of which the examiner might find it necessary to assist the examinee. The *Stanford Revision of the Binet Test of Intelligence* is essentially a highly controlled interview in which the examiner must meticulously follow prescribed procedures. The method of administering the test requires personal interview with each child tested.

The personal interview technique often is more revealing than other techniques of obtaining information. The interpersonal relationship between examiner and examinee may elicit incidental revelations from the examinee which suggest unexplored aspects of a problem. The examiner may assist the examinee in interpreting topics upon which information is desired. If proper rapport characterizes the interview, the examinee may reveal himself more completely than he would in making his statements in writing. Although it is a canon of research that only data definitely relevant to the purposes of an investigation should be collected, investigational by-products obtained often prove unexpectedly significant. The personal interview sometimes permits free association about questions asked, in much the same way that an essay item enables the learner to display breadth of knowledge to a

greater extent than is possible when an objective item is used.

In many cases the interview technique is implemented by use of stenographic records and phonographic and photographic recordings. From material gathered by such devices it is possible not only to record but later to verify certain responses. It is often difficult for an examiner to maintain the desired degree of rapport between the examinee and himself if he must continuously make written records during the interview. Yet to record information following the interview obliges him to depend upon memory which becomes less dependable with lapse of time. Use of rating devices or score cards constitute variant aids in the control of the interview. Their function is similar to that of an outlined procedure defined in the form of questions.

The interview technique was the major means of obtaining data in an investigation of children's[1] explanations of natural phenomena. Individual interviews were held with members of four groups of children (kindergarten, and grades 2, 4, and 6) concerning their explanation of various types of natural phenomena. Answers were analyzed in terms of a number of subcategories under the main classifications of *physical, nonphysical,* and *failure to explain*. All responses to each question were tabulated on a separate tally sheet for each grade, with separate listings for each part when a response included different ideas. Responses were recorded during the interviews. Quantification of data was made possible by appropriate planning of classifications.

An abridged sample[2] of a typical question and their various explanations follow. Certain types of the subclassifications are shown under the categories: *physical, nonphysical,* and *failure to explain.*

(The topic is: *Stream Flow*. The children were asked to explain why the water in a creek with which they were familiar flowed.)

PHYSICAL

Complete—Because it's downhill (or sloping to lake). There's falls (or rapids) all the way along.

Partial—Because there's water coming into it. 'Cause rain makes it.

Simple Phenomenalism—'Cause the waves make it. It has to or it'd flood. Because the wind (or air) blows (or pushes) it. There's rocks to it. It swims on a stone. Because it's wet.

[1] M. E. Oakes, *Children's Explanations of Natural Phenomena* (Teachers College Contributions to Education, No. 926, 1947, 29, 30).

[2] M. E. Oakes, *op. cit.*

Human—They dig it that way. They got a hole—water come up. 'Cause a pump makes it. It comes out a' iron pipes.

Reversal—Boats make it go. The mill wheel pushes it.

NONPHYSICAL

Animistic—The water wants people to get in. 'Cause the boats want it to. 'Cause the cars want to go by.

Magical—'Cause it comes from the ground. Under the ground there's water.

Religious—God makes it move along. Jesus goes into the water and pushes it.

Teleological—'Cause so the fish can swim. Because it gives other lakes some water. So it can go into other creeks.

Providential—'Cause so we can get some water. Because it has to go along, so when they milk the cows they can wash the milk pails. Because to make boats go along. 'Cause big kids need it to swim.

FAILURE TO EXPLAIN

Obvious—The water just goes by itself. 'Cause it's water.

Restatement—Because it goes following all the rest of the creek. It goes into another lake. 'Cause it comes from another place, probably. 'Cause there's a current in it.

Irrelevant—'Cause it's skinny.

"I don't know"—I never thought of that. Nobody told me that. In winter water isn't in creeks, 'cause it doesn't rain as much as in summer. I don't know—probably because the things make it.

Physical interpretations were most numerous, increasing from the lower to the higher grades, with a corresponding decrease in explanations of a nonphysical character. Differences in types of explanation given by children with high and low IQs were not great, although bright children proposed physical interpretations more frequently than the less bright. Compared with data from a supplementary study of explanations given by college teachers in nonscience courses, those of children showed no essential differences in the kind of thinking involved. No evidence was found that there are definite stages in a child's thinking which are characteristic at given ages. Replies seem influenced by the nature of the problem, its wording, and the child's experiential background and command of vocabulary more than by any so-called "mental structure" at a given age.

The questionnaire. The term "questionnaire" generally refers to a systematic compilation of questions that are submitted to a sampling of population from which information is desired. Usually when a questionnaire survey is conducted an addressed

stamped envelope is supplied. Follow-up letters referring to the significance of the study often result in increasing the percentage of returns. The information required usually involves statements relating to (1) fact or (2) opinion. The questionnaire is generally regarded as more dependable when used to obtain statements of fact. Proper selection of respondents becomes especially important when opinions or judgments are sought.

The questionnaire technique is similar to that of the controlled interview. Its use may be desirable when personal interviews would be costly or difficult to arrange. The questionnaire makes possible contact with a large number of persons and also with many who could not otherwise be reached. Information of a personal nature often may be obtained more readily by means of questionnaires, especially if the respondent is permitted to omit signature or if specifically assured that his replies will be regarded as confidential.

The technique of collecting data by means of questionnaires is frequently ineffective for purposes of accurate investigation because of (1) improper formulation of questions, (2) improper sampling, (3) inadequate returns, and (4) failure to select respondents who are capable and willing to co-operate. Effectiveness of the questionnaire may be increased by requiring information which can be supplied with minimum difficulty. If detailed writing or verbal description is required, rapid and objective tabulation of results is difficult. Short-answer items often will elicit desired information in readily usable form.

Inasmuch as analysis of questionnaire data is dependent upon classification of information, the questionnaire should be designed so as to facilitate such classification. The items control the extent to which the respondent must make discriminations, but the range of possible responses should be completely provided for. "Catch-all" categories suggest that the significance of possible responses has been inadequately explored. Vague terms such as "often," "much," "to any extent," or "usually," as well as negative terms such as "no" or "never," especially in "yes-no" questions, tend to be confusing.

The apparent ease of planning and using a questionnaire tend to make it appealing to novices in research. This uncritical attitude has resulted in many surveys conducted by means of poorly prepared questions.

For research purposes necessitating precise information, the questionnaire must be more carefully constructed than one intended primarily for the status-survey type of investigation.

Methods of determining reliability and validity of a questionnaire are often unwieldy and unsatisfactory. More dependable results are ensured by (1) minimizing the effects of certain conditions which impair reliability or validity and (2) designing the items with a view to quantification of results. The following criteria may be used as a guide in constructing questionnaires:

1. Objectivity in meaning and scoring should be sought.
   a. The questions should be formulated so as to enable the individual to
      (1) supply information under discrete categories
      (2) express specific points of view.
   b. It should be possible to translate replies into quantitative expressions of absolute or relative values which may be described by statistical techniques.
   c. The plan for quantitative treatment should be sufficiently adequate to permit
      (1) question-to-question comparisons
      (2) comparison of findings with results of other investigations.
2. The appropriateness of recall and recognition items for eliciting information should be determined.
3. Opportunity should be given the respondent to include supplementary or explanatory information. In some instances the respondent should be encouraged to provide fine distinctions or to include qualifications or reservations.
4. The questionnaire as a research instrument should be sharply focused upon specific purposes and be analytical in nature.

Every precaution should be taken when formulating the wording of items soliciting expression of opinion that such wording is "neutral" and does not convey the impression that any opinion is favored by the investigator or is to be favored by the respondent over any other opinion. References to attitude-stereotypes should be avoided. An item mentioning by name a country with which we might be at war or not in cordial relationship for example might suggest that an unfavorable opinion should be given. The respondent may be expected to report his opinion more thoughtfully if asked to state his point of view concerning practices, policies, or theories presented without mention of the country where they might exist.

If possible, a tentative form of the questionnaire should be empirically evaluated with an experimental group of competent persons in order to detect faulty construction. Such actual trial affords a basis for revision and elimination of defects and for making the definitive edition more efficient.

The following questionnaire is reproduced in part as an illustration of one which was used to obtain a large amount of useful information. The complete form identified the author of the questionnaire and the sponsoring organization.

The Bureau of Educational Research, in co-operation with the Bureau of High School Visitation, is conducting a survey of the testing practices in representative secondary schools of the state, and we request you to fill out the attached form. (*Estimated time required: 15 min.*).

Inasmuch as testing practice varies with different courses and grade levels, we ask you to indicate the courses you teach. If you teach more than one subject, select the *one* which you believe most representative of your general testing practices. And then answer the questions STRICTLY IN ACCORDANCE WITH YOUR PRACTICE IN THAT COURSE ALONE.

The investigation is concerned with *written tests* only. Some questions invite your criticism and opinions. Please feel free to express the best that your experience has indicated. Don't worry if your opinions run counter to "textbook" advice.

1. What is the usual make-up of your examinations? (Check one).
   ______Essay questions only
   ______Essay with 1, 2, or 3 types of objective test
   ______Essay with 4 or more objective tests
   ______One type objective only
   ______2–3 objective tests only
   ______4 or more objective tests only

2. Rank by number (1,2,3, etc.) the kinds of test you employ most frequently—1 being the most frequent. Draw a line through those you have not used.
   ______True-false
   ______Multiple choice
   ______Completion
   ______Problems
   ______Matching
   ______Essay type
   ______Other (specify)

3. How often do you give written tests? (Check one).
   ______Daily
   ______Twice weekly
   ______Weekly
   ______Monthly
   ______Other (state approx. time)
   (Number of minutes devoted to each such test: ____________)

4. After you have graded test papers, do you require students to correct their wrong answers?
   Yes______; No______.

5. When do you usually return test papers to students? (Check one).
   ________Next class period ________Tests not returned
   ________2–3 periods later ________Other (specify).
   ________One week later

This questionnaire was addressed to teachers to ascertain trends in actual testing practices. In connection with the same investigation, another questionnaire [1] was used to obtain student opinion concerning testing practices. A portion of the questionnaire is reproduced as follows:

TO THE STUDENT (Grades 10, 11, 12 only):

This survey is being made to obtain information necessary for research work at the Bureau of Educational Research.

You are asked to give each question careful consideration and to answer it to the best of your knowledge. The questions, as you see, concern your opinions and preferences regarding testing practices in your school.

Although we request certain information about your status in school, you are asked NOT to put your name on this questionnaire. We expect your answers to be of real help; so feel free to state your honest opinions.

1. In your opinion, how should tests count in making up the final grade for a course? (Check one.)
   ________Only exams which mark the end of units-of-study should count.
   ________Only mid-term and final exams should count.
   ________Exams should count more than quizzes and assignments, but all should count.
   ________The grade should be based on the final exam alone.

2. Which do you prefer? (Check one.)
   ________Learning your grade but not getting back the test itself.
   ________Getting back the paper with the grade on it but nothing else.
   ________Getting back the paper with the grade and corrections on it.

3. Which of the following types of test do you prefer?
   ________Those in which you must depend solely on memory for exact facts and figures.
   ________Those in which you must use facts and information in solving new problems.

[1] W. Bender and others, "What high-school students think about teacher-made examinations," *J. Ed. Res.*, 1949, *43*: 58–65.

4. After an exam has been graded, should the students be required to correct their errors and return these corrections to the teacher?
   Yes________; No________.

5. Of the following testing plans, which one do you think would give you the best chance of showing what you know?
   ________Daily tests . . . 5 or 10 min. long.
   ________Weekly tests lasting 20–30 min.
   ________Monthly tests lasting one full class period.
   ________Mid-terms and finals only.

Quantification of data obtained by the typical questionnaire is generally achieved through tabulation and counting. Refinement of results in tabular form in totals, percentages, or averages, and calculation of coefficients of correlation often is made in order to suggest probability of relationships among data. Means and medians are frequently computed. The data are expressed quantitatively on the basis of the number of persons whose replies are tabulated under the several categories of the questionnaire. The method is similar to that used in studies based upon interviews or observation. Emphasis is upon the facts or opinions revealed; foot-rule or ranking methods are not required in order to establish a basis for comparing individuals. Independent categories of information, however, are necessary for extensive treatment of results.

Inventories. The term "inventory" refers to a pencil-and-paper type of instrument designed to reveal an individual's typical behavior in connection with various traits of personality, attitudes, and interests. By means of inventories various attributes of individuals may be quantified. For example, we quantify the extent to which a person possesses a tendency to dominate in social situations. The instrument used for measuring a tendency to dominate is designed to yield for this trait different numerical scores.

The purpose of an inventory is to determine the extent of a trait. We may wish to know, for example, whether an individual possesses an extremely strong tendency to dominate in social situations, an extremely strong tendency to be dominated, or a tendency to reveal dominant and submissive behavior in mixed amounts.

The desirability of the types of behavior involved is beyond the scope of inventories. There may be social prejudices for and against persons possessing certain combinations of personality traits. Social prestige may be attached to some occupations to a greater extent than to others. Certain defense mechanisms are presumably in operation, for example, when a practical workman

scorns the advice of a technically trained expert as being "impractical theory." Every evaluation of desirability is probably influenced by the attitude of the person who makes the evaluation. If the evaluation of desirability is to be made, it must be made by other methods.

The practice in quantifying personality is to assume that there are certain fairly discrete patterns or traits and that an individual's typical modes of adjustment are revealed in his consistency of behavior. A person may be considered introverted-extroverted, unsocial-social, ascendant-submissive, shy-bold, or carefree-worrying. These are samples of the extremely large number of trait names which have been used at one time or another in the study of personality. The extent to which a "personality-total" may be completely described by means of a relatively small number of "factors" is a moot question in current study of personality.

A similar situation is observed in inventories of interests. An individual's vocational interests may manifest a certain degree of clustering about generalized patterns. Many of his behaviors are believed, for example, to be associated with possession of "mechanical ability"; concentration of positive behavior in this direction may afford some assurance of success in certain occupations. It may be difficult to agree upon the relative importance of different kinds of behavior which indicate possession of mechanical ability. In one instance, interests of persons successful in certain vocations were used as a basis for standardizing a vocational interest inventory. Given a reasonably valid formulation of types of specific behavior, it is possible to quantify the degree of resemblance between the individual's pattern and a standard pattern.

Quantification of a trait of personality begins at a point at which lists of specific manifestations of behavior are formulated and values assigned to them. Such lists are usually a result of intensive analysis and experimentation, with continuous effort to achieve a high degree of validity. Research workers often use standardized instruments which have been in process of development during extended periods of time.

An approach frequently used in presenting the inventory to an individual is to require a "yes" or "no" answer, according to whether he believes the instance of behavior is characteristic of his habitual performance. He is sometimes directed to include a "?" or "0" if he feels unable to make a decision. With reference to some trait of personality, an individual may be asked whether he would "enjoy knowing that he is to be guest of honor at a dinner." Even though he may not have previously considered how

he would react in such a situation, his emotional and intellectual background holds an answer for him; and he can usually decide without hesitation whether he *would* or *would not* enjoy such an honor. It is of no importance whether his answer is a formulated belief or a spontaneous reconstruction of experience. In fact, a studied reaction on his part is less desirable than one in which he replies without reflection in response to his attitudes.

Some instruments make use of multiple choice items which have the effect of "forcing" a choice even though no choice is completely acceptable. There is a tendency to word the items in such a way as to prevent the individual from successfully guessing the aspects of personality to which any instance of behavior alludes. The inventory type of instrument may become less reliable if it is possible for the individual to vote for the aspect of personality personally acceptable to him rather than to reveal the actual nature of his behavior. It may be difficult to obtain a truthful answer if an item requires confessing socially unacceptable behavior.

In many types of instrument, a specific item of behavior is designed to permit multiple inferences for each of several traits by use of different keys for scoring. Score values may be weighted differently according to the trait being measured or applied to different directional aspects of traits. A "yes" underlined or checked for a given behavior may be assigned + 3 for 'introversion' and − 1 for "self-sufficiency." The same "yes" may be assigned "0" for "emotional stability" and − 4 for "neurotic tendency." Weighted values are determined in accordance with the extent to which certain behavior is a significant indicator for the traits studied. Intermingling of items and provision for multiple inferences are used as bases for preventing the individual from "working" the test.

Traits of personality are generally considered as paired-opposites. In a trait designation such as "introversion-extroversion" either of the terms may be regarded as being opposite. Descriptive behaviors, however, are usually stated in positive form regardless of the direction to which they have reference, since their significance may be represented by positive or negative score values. As a rule a type of specific behavior is used only once in an inventory and is seldom in the form of positive and negative aspects of the same action.

Current belief is that relatively few individuals are typically introverts or extroverts in all situations and that behavior with respect to these traits tends to be specific to the situation. It has been similarly claimed that honesty-dishonesty as a character trait

should be regarded as specific to the ethical situation. Individuals may be honest in some situations and dishonest in others.

Positive and negative numbers afford a convenient basis for expressing quantity with respect to a trait of personality. The continuum of values upon which to indicate an individual's position with respect to a trait may be devised to range from high negative numbers at one extreme to high positive numbers at an opposite extreme. Such scores may be obtained as -173, -145, -14, 0, $+5$, $+75$, or $+158$. A score of "0" may be interpreted as meaning that an individual is not inclined toward either extreme of an opposed pair of traits, provided zero is the actual midpoint of all possible scores.

The index of value for a trait consists of the total of possible points, added algebraically in the case of positive and negative numbers. If there are no weighted values, a score for a trait may be the number of instances of behavior characteristic of one aspect of a trait minus the number of instances of contrary behavior. The plan of many instruments is in some respects similar to that of a test consisting of true-false items.

As was emphasized earlier in this chapter, no number has meaning without a frame of reference. In measuring a table by means of a scale, it is important to know whether an obtained index of length is interpretable as inches or centimeters. In the case of an index of value for a trait of personality, a score is meaningless unless it can be interpreted (1) by some knowledge of the nature of the trait and (2) by the range of possible scores obtainable. There are no standard units for quantifying an "amount" in connection with a trait.

Quantification rests upon the basis of the extent of significant behavior. The more frequently the individual's characteristic behavior related to an aspect of a trait occurs, the greater will be his score. An inventory is initially scored by a method which is essentially that of counting or totaling the values assigned for each response with reference to the trait to which behavior described in the item relates. In the interpretation of scores reference is made to a standardized scale for the instrument used. In a sense, interpretation is made upon the basis of ranking. Since, however, the items of a personality inventory are not "correct" or "incorrect" responses, a percentile or standard score indicative of rank order position is derived principally for statistical purposes. The purposes of an inventory otherwise may be fully satisfied by a simple description of rank position and its significance in terms of the trait. A score for "introversion-extroversion" may be given

a letter designation, such as "A" for tendency toward ambiversion, "E" for tendency toward extroversion, or "I" for tendency toward introversion. Since a score suggesting a "personality-total" serves no practical purpose, such scores are seldom computed.

Rating techniques. "Rating" is a term applied to expression of opinion or judgment regarding some situation, object, or character. An individual for example may be rated with respect to various aspects of efficiency. Opinions are usually expressed on a scale of values. Rating techniques are devices by which such judgments may be quantified.

Suppose it is desirable to measure an individual's "loyalty." There are no footrule methods which apply to such a trait. Definition of loyalty in terms of its operational characteristics must be the initial step in the process. After a large number of specific characteristics or traits bearing upon loyalty have been formulated, ratings may be made from several points of view and combined into a single score. Yet, the index obtained would not represent a certain quantity of loyalty.

An individual's rating, for example, may be 20 on a given series of ratings, of which the possible attainable score is 25. This fact enables the loyalty of that individual to be compared with that of other individuals who have been subjected to the same rating procedure. When 25 is a maximum score, it is of value to know that an individual has attained 20 points. It is probable that his loyalty is more intense than that of a person whose ratings total only 10 points. The principal challenge to the value of the information is whether the rating scale measures consistently and whether it is a valid measure of loyalty.

The effectiveness of measurement by rating methods requires specificity and comprehensiveness of definition. Measurement of a quality such as "efficiency" is difficult because it may be expressed in many types of behavior. For a given situation, an individual's efficiency may be judged by such criteria as: (1) Is he prompt in beginning his daily work? (2) Does he follow a plan for grouping related tasks? and (3) Does he use his time exclusively for allotted tasks?

In the measurement of "efficiency" each proposition or question involves an assumption, subject to verification, that individuals are "efficient or "inefficient" according to the extent of truth contained in operational statements. Rating scales of efficiency formulated by two persons acting independently may yield different scores for the same individuals. If rating scales are to yield

dependable data it is necessary that individuals agree on the classification and meaning of whatever is being measured.

The ability of the rater to discriminate, imposes a limit upon the number of degrees of discrimination which are effective. The limit generally agreed upon is seven in cases in which a scale is designed for responses ranging from absence of a characteristic at one extreme to its maximum presence at the other. In such a scale, the "average" or theoretical mean appears midway.

Rating techniques are sometimes used in instruments which are not rating scales by strict definition. Items for which ratings are made appear with letters or numbers in serial order, or with words descriptive of a quality of rating, such as "excellent," or "fair," and "poor." Such a form might be the following:

YOU ARE RATING *Henry Jones* AS A CANDIDATE FOR APPOINTMENT AS HEAD BOY OF TAPPAN SCHOOL

PLEASE CIRCLE THE LETTER BEST DESCRIBING THE RESULTS OF YOUR OBSERVATION OF HIM. "A" INDICATES AN EXTREMELY FAVORABLE OPINION OF HIM, AND "G" AN EXTREMELY UNFAVORABLE OPINION.

A B C D E F G—He is courteous at all times.
A B C D E F G—He is neat in personal appearance.
A B C D E F G—He has a reputation for clean sportsmanship.
A B C D E F G—He is a positive influence toward good citizenship.

Quantification is effected by transmuting letter designations or verbal characterizations into numbers, and by computing a total score. The total score is a number representing an empirical index of value interpretable in terms of the rating device used. The rating device serves as a measuring technique of making comparisons among different individuals for a definite purpose on the basis of predetermined criteria. A rating device in this instance serves as a means of emphasizing specific qualifications and minimizing any inclination to consider only general popularity. Even though the instrument were not highly valid, its use would tend to reduce subjectivity of judgment.

Rating devices are usually constructed upon the assumption that characteristics generally rated will take the distribution form of a "normal probability" curve. The score of an individual is more likely to occur within the "average" band of the range of scores than at either extreme. This situation may be illustrated by placing the values of ten salesmen in rank order of merit: 1,2,3,4, 5,6,7,8,9,10. The typical salesman is more likely to be ranked 4,5,

6,7, than in other positions in the series. It is also easier for him to improve in efficiency and advance, say, from 6 to 7, than from 9 to 10.

By strict definition the rating scale is characterized by the provision of a line whereby the rater can mark the point at which he wishes to record his rating. Such a device is often called a *graphic* scale. Values may be set at appropriate points. It is similar to a footrule, but with two exceptions. One is that the units of value assigned are chosen and selected by the designer of the scale. The other is that there is no assurance that qualities represented are equally spaced in amount.

A common form consists of a horizontal line marked into equally spaced intervals. A five-space scale may bear for each interval-numerical designation (1,2,3,4,5), verbal designations ("never" to "always"), or both numerical and verbal designations. If the intervals are not designated by number, the investigator transmutes the letters or verbal terms into numbers when scoring. Odd numbers or intervals are most commonly used, since such a plan permits marking a midpoint as neutral or average rating. A graphic scale such as one that includes a line presented without demarcation into intervals, is sometimes used. The rater is then required to estimate visually a distance from the end of the line in comparison with the total length of the line. The rater merely designates his rating by making a point on the line.

Absence of marked or defined intervals is theoretically in accord with the fact that many human traits do not vary in amount from one individual to another by discrete steps or intervals. "Efficiency" may vary in so infinitely small amounts that highly precise determination of the point at which an individual changes in "efficiency" is impossible. Discrete intervals, when represented on a scale, are only coarse measures of a trait used for convenience. The unbroken line creates the illusion that fine discrimination is both possible and desirable and suggests to the rater a need for care and accuracy. In scoring ratings, however, the investigator must convert the rater's marks into mathematical units which may be quantitatively expressed.

A typical example of a rating scale is the *22-Trait Personality Rating Scale* prepared by Tschechtelin.[1] This scale was prepared primarily to study self-rating by elementary-school children as

[1] S. M. A. Tschechtelin, "Self-appraisal of children," *J. Ed. Res.*, 1945, *39:* 25–32. See also: S. M. A. Tschechtelin, "A 22-trait personality scale," *J. Psych.*, 1944, *18:* 3–8; "Factor analysis of children's personality rating scale," *J. Psych.*, 1944, *18:* 197–200; and "Children's rating of associates," *J. Exp. Ed.*, 1944, *13:* 20–22.

compared with ratings made of them by teachers. Ratings in respect to 22 *aspects of personality* were made using a line representing a continuum and affording opportunity for ten degrees of evaluation. The line was drawn on the blackboard in all classrooms where the scale was used, and the significance of its numbers ranging from 1 to 10 was explained orally. The form of the scale follows:

'	'	'	'	'	'	'	'	'	'
1	2	3	4	5	6	7	8	9	10

Each rater was asked when rating someone, to place in a column at the right of the questions the number indicating his rating. Some of the items were:

1. Is he "peppy" and full of life?
2. How bright or intelligent is he?
3. How friendly and sociable is he?
4. Is he restless or nervous?
5. Is he popular with other children?

These items were directed toward such aspects of personality as "pep," intelligence, sociability, nervousness, and popularity. The study revealed a consistent tendency for girls in grades four to eight to rate girls more liberally than boys. The study also showed a tendency for each girl to rate herself more liberally than she was rated *(a)* by boys in her class, *(b)* by other girls, or *(c)* by her teacher.

Rating techniques afford a valuable means of testing the validity of many objective instruments, such as pencil-and-paper inventories of personality traits. Rating scales are frequently used in determining efficiency of employees, their qualifications for various kinds of work, and in many similar situations in which canvassing opinion is the only source of dependable information. The technique has also been helpful in the appraisal of school buildings and school systems.

Rank-order scaling. Comparison of members of a group with respect to certain qualities commonly shared in varying degrees is sometimes made by means of rank-order scaling. Names of individuals are placed in serial order with respect to each of the variable qualities. Members of a group might, for example, be ranked according to the degree of leadership manifested by each member. Rank-order scaling is useful as a means of dealing quantitatively with qualities or attributes which have not been differentiated clearly.

There are no numerical units in a rank-order scale other than

the numbers indicating the serial position of each member. Since the concept of normal probability must be applied in considering the distribution of human traits, rank-order intervals may not be expected to be equidistant. There is likely to be a greater number of "average" than "excellent" or "inferior" individuals; and at midrange individual differences are more difficult to detect and to assign to a specific rank order. Individuals at either extreme of the scale are more likely to be ranked reliably than those occupying central positions.

Certain general or highly elusive traits, such as honesty, sincerity, tactfulness, or interest in one's work, may be efficiently dealt with by ranking individuals on an over-all basis, without a high degree of trait analysis. The technique is especially applicable in evaluating products. Advertisements for example, have been rated for their attention value, persuasiveness, and memory value. Ranking techniques are frequently applied in estimating the relative merits of essay items, themes, or term papers. In a certain instance, the relative value of Oriental rugs was determined by computing pooled judgments of rank values.

Rank-order scaling becomes increasingly reliable up to a certain point as the number of raters is increased. Their judgments may be pooled by computing a mean or median rank for each quality in respect to which members of a group are rated. The "average" rank determined still indicates, however, serial order and not scale value.[1]

The man-to-man technique. The *man-to-man* technique involves both ranking and rating at various points in the complete technique. A number of persons (usually three or five) known to those who are to rate is initially selected to serve as living descriptions of the highest, lowest, and intermediate degrees of some trait or characteristic. The rater ranks the individual among a series of persons selected for comparison with respect to each trait or char-

[1] It is possible by means of simple calculations to convert rank order assignments into units of amount or "scores" upon a linear scale. Such conversion is not advisable unless it is possible to assume a normal probability distribution of the frequency for the trait in which ranking has been made. The following formula (from H. Garrett, *Statistics in Psychology and Education,* N. Y., Longmans, 1937, p. 169.), which converts ranks into "per cent" positions, is first used:

$$\text{Per cent position} = \frac{100\ (r - .5),}{N}$$

where R is the rank in the series and N is the number of individuals ranked. One may then consult a table (Garrett: p. 171, Table 28, TRANSMUTATION OF ORDERS OF MERIT INTO UNITS OF AMOUNT OR "SCORES"), which will yield an equivalent score upon a scale of 10 points.

acteristic considered. This type of instrument was initially designed in 1917 for use as the *Army Rating Scale,* to evaluate personnel under five headings: physical qualities, intelligence, leadership, personal qualities, and general value to the service.

Guthrie[1] used this technique in a procedure for evaluating faculty members for promotion and increase in salary at the University of Washington. Evaluation is made by a secret committee which does not meet but which reports individually to a personnel executive. The candidate is first asked to supply information about himself and to bring the bibliography of his writings up to date. He is invited to make four nominations for his evaluating committee. The committee is appointed, three members from the candidate's own department, two from allied departments, the executive of the candidate's department; and the dean of the college, if he considers himself qualified to serve as a rater. Ratings are made on a man-to-man basis as may be noted in the following condensation of the materials furnished each rater:

> To indicate your opinion, first fill in the blanks on the next page, in the order of what you believe to be their value to the university, with the names of five faculty members (without regard to rank) in the candidate's department or in closely related departments. Choose one who is *outstanding,* one who is *superior,* and one who is *only fair,* and one who is *only of slight value.* Write the names in the order of their merit from best to poor.
>
> . . . .
>
> For each of the following items consider where, if inserted in the list of five, the candidate being considered belongs. His name when inserted in this place will make a total of six names.
>
> Encircle after each item the number (from 1 to 6) which indicates the candidate's position for that item.
>
> 1. teaching effectiveness 1 2 3 4 5 6
> 2. contribution to his field through research and publication
> 3. value of his departmental and campus activities (other than teaching and research)
> 4. value to community and state
> 5. ability to cooperate with the members of his department
> 6. knowledge of his subject
> 7. general knowledge and range of interest

[1] E. R. Guthrie, "The evaluation of teaching," *Educational Record,* 1949, *30:* 109–115.

8. rate of professional growth (recent)
9. recognition by others in his profession.

The product scale. Brief mention should be made of a technique of making rank order comparisons known as a *product scale*. Such scales have been used as devices for evaluating handwriting, lettering, and English composition. The device consists of a series of samples of the product under consideration arranged in order of merit on the basis of the consensus of raters' opinions concerning the degree of merit shown in each sample. Scores are attached to each sample indicating relative degrees of merit. A specimen of a pupil's handwriting for example may be compared with each sample.

Usefulness of the method is highly restricted to quantification of types of ability which may be displayed in the form of attempts to satisfy standard requirements. The method is applicable in a slightly different form in making comparisons of art products, when a standard task may be required. All pupils may be required, for example, to paint in water color a picture of some object, such as a vase. Products of such activity may be compared on a fairly uniform rank order basis since each may be referred to a standard sample, consisting of the vase itself or a picture painted by the instructor.

Quantification in the case of a product scale involves certain aspects of rank-order scaling. The actual ranking of products is completed before the product scale is used to evaluate the merit of a given specimen. A product scale also possesses characteristics of a footrule in that differences between successive standard samples may be considered equidistant in amount of merit. Assignment to a position on the scale of values involves an act of rating. If this rating is accurately made in the case of a specimen of the product, the process of using a product scale results in achievement of a certain amount of objectivity, since the scale may be standardized upon observable characteristics of the product desired.

The method of paired comparison. Among the rating techniques that have proved valuable in measuring attitudes is that of *paired-comparisons* instead of rank-order scale. The following example illustrates the method of paired comparisons in determining attitudes toward nationalities: [1]

[1] R. Stagner, "Fascist attitudes; their determining conditions," *J. Soc. Psych.* 1936, 1: 438–454.

This is a study of attitudes toward nationalities. You are asked to underline the one nationality that you would rather associate with. For example, the first pair is Englishman-Norwegian.

If, in general you prefer to associate with Englishmen rather than with Norwegians underline Englishman. If you prefer, in general, to associate with Norwegians underline Norwegian. If you find it difficult to decide for any pair, be sure to underline one of them anyway. If two nationalities are about equally well liked they will have about the same number of underlinings in all of the papers. Be sure to underline one of each pair, even if you have to guess.

Englishman-Norwegian	Norwegian-Irishman	Greek-Pole
Swede-Belgian	Swede-Russian	German-Austrian

Lists of words are sometimes used. The individual is asked to check according to the pleasant or unpleasant connotation of each word to him. The success of this method depends upon the intensity of the attitude under consideration and the individual's co-operation in registering his immediate reaction.

An adaptation of this technique has been employed by Lundberg [1] in rating a group of students on one factor. The names of students are listed on two sheets of paper. A stuffed paper folder is cut making a slot large enough to expose the name of one person at any time. This person is then compared in turn with every other person on the list by making a check after the better one in each case. Then the next name on the list is exposed in the slot and compared with everyone on the list. This procedure is continued until each person has been compared with every other person on the list. It is possible by this means to rank each individual on the basis of the number of checks received. Inasmuch as both lists are checked according to this method it is possible to obtain a measure of the consistency of each rater.

Rating scales and rank order scaling compared. The method of quantification used in a rating device is essentially that of rank order. The observer determines the rank order position upon a scale of values at which various traits possessed by an individual, for example, are to be located. Such allocation is upon a "more than" or "less than" basis. The major difference between rating techniques and rank-order scaling lies in the points of emphasis. A rating scale usually stresses the rank order values of many traits or qualities of individuals, and comparisons are made on the basis of score values. In rank-order scaling, emphasis centers upon direct

[1] Donald E. Lundberg, "A simple rating device," *Personnel Journal*, 1917, *25*, 267–270.

comparison of individuals upon the basis of various qualities, each considered singly.

Both techniques place a premium upon the qualifications of raters and upon rating conditions, since considerable subjective judgment is involved. For this reason, success of either technique depends upon proper selection of raters. Since no amount of statistical manipulation can compensate for errors and inaccuracies which are especially likely to occur during the subjective phases of rating and ranking methods, certain suggestions are included which may improve the quality of the initial data.

1) *The raters should be carefully selected.* Their selection should be based not only upon competence but upon other considerations. Individuals who are usually confident in their personal decisions are likely to rate accurately and rapidly. They should be acquainted with the individuals whom they are to rate and be conscientious. Although it is usually desirable to have several independent ratings by different raters, the quality and training of the raters is fully as important.

2) *Raters should be trained for the particular job.* The investigator should make clear the expected interpretation of all wording pertaining to the variables which are to be used in rating. He should explain any ambiguities by suggesting synonyms or by explaining the operational aspects of certain behavior. Raters should know precisely the basis upon which a rating is to be made. Significance of the scale intervals should be explained, with instructions to avoid clustering ratings too conservatively about an average point or too literally at extreme points.

3) *Overt behavior can be more readily rated than that which is hidden.* Certain behavior, such as "leadership," is more accessible to observation than inner behavior such as one's "appreciation of art." Introverted individuals are usually more difficult to rate than individuals who express themselves openly and freely. Raters tend to rate with greater accuracy those individuals with whom they share common characteristics. It is generally believed that a rater can rate more accurately a member of his own sex.

4) *It is important to resist the "halo effect" of some strongly dominating characteristic of an individual.* A teacher's pleasant manners may lead a rater to believe uncritically that the teacher maintains good control over his class or marks tests fairly. The halo effect may be minimized by specifically cautioning raters against such danger and by emphasizing the importance of rating the individual in terms of the variables as specifically described.

It is also possible to vary the descriptions of each variable in such a way as to make a fresh independent decision necessary.

5) *Many raters are inclined to prefer favorable to unfavorable statements.* Such inclination presumably stems from a tendency toward forbearance and generosity which develop as part of an individual's ethical culture, causing him to regard low ratings with feelings of guilt. This tendency may be counteracted by eliminating all wording which makes a rating appear to have a detractive or derogatory implication. Social acceptance of certain personal values tend to introduce systematic error. "Leadership" is regarded as a highly desirable quality, whereas "lack of leadership" is associated with an inferior status. The tendency to rate favorable aspects of individuals highly is reduced in ranking techniques in which all grades of merit must be occupied by someone. As Allport [1] remarks, "Even a choir of angels may have its least favored members."

Testing Situations. Our interest in testing situations is expressed primarily in considering their use in quantifying achievement of individuals who have been under the influence of instruction. Our discussion, however, is applicable to tests of ability such as intelligence and aptitude, in respect to the principles of quantification involved.

The achievement test constitutes the principal instrument used in measuring the extent to which learning has occurred, as well as being at the same time a means of facilitating learning. The broad scope of desirable pupil achievement is explicitly defined in many formulations of instructional objectives. Possession of information in increasing amount is an important instructional aim. Display of reinstated information however, furnishes no evidence of the extent to which such information has been organized among the results of the learner's earlier experiences, the extent to which such information facilitates his future thinking, or the extent to which he has mentally associated such knowledge with other information in his possession.

It is sometimes assumed that achievement exists in determinable or finite amounts. Even in the case of informational content, it is rarely possible to determine quantitatively the amount of information which the learner possesses. Learning material is not capable of being measured by any fundamental process of meas-

[1] G. W. Allport, *Personality: A Psychological Interpretation* (New York, H. Holt, 1937, 445).

urement. Measurement based on the mastery of certain chapters of a textbook or of specific items of information is not true quantification but rather qualitative description of amount. The practice of regarding learning material as units of measurement for purposes of quantifying achievement can result only in rough estimates.

In the light of these considerations, the basis of a percentage scale of measurement may be misleading. It is proper to interpret a per cent grade of 75 as indicating that 75 per cent of the responses made upon a given test are correct. It is not proper, however, to interpret 75 per cent as indicative of an amount of absolute achievement. To do so suggests that 100 per cent represents a standard or criterion amount of learning material. An arbitrary hurdle is erected for the learner, which might be expressed by saying: "The amount of material for which you are responsible consists of five chapters of the textbook, and you are to be able to state any fact presented in these chapters." In such a case, quantification has been achieved, but it is far from valid in view of the real nature of learning.

If objectives other than ability to reinstate information are emphasized, a study of what is involved in their attainment may help to clarify the problem of quantification. The ability to apply information results in identifying a situation in which given information is applicable. If a student makes 75 applications correctly among 100 attempted, the only quantitative fact of which there can be certainty (upon the basis of this evidence alone) is the percentage of his success in the situation arbitrarily set for him.

Achievement a relative term. We approach the problem of quantification objectively when we think of achievement as a relative term and as demonstration of behavior that must be evaluated comparatively. No relative importance may be attached to the fact that a pupil has been able to solve 10 problems in two minutes. But let us suppose that this task has been attempted by others and can be performed by only one individual. There is then no difficulty in assigning rank order value to the performance. The numbers in the phrase "10 problems in two minutes" express quantity; and the action may constitute a successful event in isolation, if such accomplishment were the pupil's goal. The numbers describe what has been done, but they do not indicate the importance of what has been done. The significant fact is that the numbers in themselves do not enable evaluation of achievement to any greater extent than was the case in which 75 situations were correctly dealt with. The significance is based on the concept of

"more than" or "less than" that of other persons. Evaluation of achievement must be on the basis of relative performance.

Evaluation requires that each individual's performance compare with the performance of the group of which he is a member or with that of comparable groups. If the individual who made 75 correct responses were in a class in which the median score was 85, his achievement could not be valued so highly as it might have been if the median score had been 45. As will be shown later, it is possible upon the basis of mean or median scores to express numerical values for the importance of individual raw scores.

A method that converts raw score values is usually available when standardized achievement tests are used. In such cases, the raw score of each individual may be interpreted by means of "norms" of performance. In this way, an individual's performance is compared with that of a relatively large group. A standardized test may be of great value as a basis for comparing achievement of a small group with typical achievement of large groups elsewhere.

If a test consisting of 120 true-false items is administered and scored by the formula (R-W), the maximum possible scores would be 120–0, or 120, with the minimum score held at 0. Scores are assigned anywhere within the range from 120 to 0; but as raw scores they indicate that an individual's score is "higher than" or "lower than" the scores of other individuals.

In tests that measure intelligence and aptitude the same principles of evaluation are involved. A raw score on an intelligence test is referable to "norms" of performance. Evaluation of the mental ability of a particular individual is made by comparing his performance with that of individuals included in the standardization group.

The aptitude test is of prognostic value in selecting individuals who are likely to succeed in adjusting themselves to the demands of some future learning situation. This is often a vocational field, requiring the performance of some task in specific or minimum amount. Performance below certain standards is unacceptable. Minimum standards may be essential in order to bar entrance of individuals who on the basis of investigations may be regarded as unlikely to succeed, for example, as lawyers or physicians.

Expressing and interpreting scores. Rank-order value of an individual's raw score of achievement stands out clearly when a frequency distribution of scores for a group is made. Whether he ranks high or low in his group may be approximately determined by inspection. Raw scores often are transmuted to some scale by

means of which their relative values may be readily interpreted.

Conversion of raw scores into percentile ranks[1] is one procedure for showing comparative achievement. Such converted scores not only continue to show the individual's rank order in the group but also indicate numerically the approximate value of his position in the group. If his percentile rank is 75, his achievement is equal to that of the highest scoring individual in 75 per cent of his group.

Two limitations should be recognized in connection with rank expressed as a percentile rank. One limitation is that the ranks at the high and low ends of the series are not accurately represented by the extremely high and low percentile ranks respectively. Middle-range scores are transmuted with sufficient accuracy for general purposes. The second limitation is that percentile scores can not be added or averaged without introducing error.

In order to facilitate further statistical treatment, raw scores are frequently transmuted into "standard scores." The calculation necessitates computation of statistical values known as the "mean" and the "standard deviation." A standard score takes into account the deviation of a given raw score from the mean score, such deviation being expressed in terms of the standard deviation. The standard score may be calculated as a "z-score." A standard score in such a form may have numerical expression ranging from −3.00 to +3.00. A z-score of .00 indicates that the individual's achievement is at the mean of achievement for his group. A z-score of +2.00 indicates that he stands at 2 standard deviations above the mean. Standard scores have an advantage over percentile ranks in that they may be added, substracted, or averaged without error just as though they were expressed in terms of pounds or inches.

The "T-score," another type of standard score, is based upon an arbitrary mean of 50 and a standard deviation of 10, after the distribution of raw scores has first been normalized. The range of such scores is from 0 to 100, although their calculation seldom results in scores at these extreme distances from the center of a distribution. T-scores are statistically comparable and may be dealt with arithmetically without introduction of error. The statistical methods briefly referred to here are applicable to the conversion or rank-order scale values obtained from achievement tests or from other sources into tangible and interpretable expressions of relative value.

[1] For elaboration of the mechanics of calculating percentile ranks, means, standard deviations, and standard scores, the reader is referred to textbooks on statistics. It is assumed that a research worker will become familiar with elementary statistics as an indispensable tool for conducting research.

For most statistical analyses, raw scores are preferable to transformed scores, since each time a transformation is made, additional assumptions are introduced. Transformations of certain kinds may be required, however, if the assumptions underlying the efficient use of a particular tool are not fulfilled. The transformations which have been introduced here may usually be used statistically if simple limitations are taken into account.

If an individual's score is quantitatively evaluated in accordance with the performance of the group within which his achievement is tested, a pupil of superior ability may earn a relatively higher score of achievement in a slow-learning group than in a fast-learning group. This problem is symptomatic of inadequate provisions of learning situations appropriate for learners of different degrees of ability.

The problem of adjusting instructional procedures in order to provide for individual differences is indirectly a problem of quantification. Such differences as are reflected in ability to achieve are likely to appear regardless of the method of evaluating achievement. Differences in the ability of different groups to achieve may be observed by comparing mean or median raw scores. The significance of a measure of central tendency, such as a mean or a median, is not, of course, complete without a measure of the variability of scores from the mean or median. Appropriate measures of variability are the standard deviation or the quartile deviation. The significance of the measure of central tendency used is greater when the "spread" of scores is within narrow limits than when it is relatively broad.

Instructional objectives are more desirably formulated in operational terms, since such expression facilitates evaluation of achievement. Test items may be constructed as samples of the various types of behavior described operationally as objectives. Demonstration of achievement should require samples of behavior representing all instructional objectives stressed. It is desirable to determine whether individuals are making progress in each objective. Some test items may be limited to reinstatement of information, whereas others may require ability to make applications of facts or principles.

The form of the test item has an important bearing upon the quantification of achievement. Careful thought should be given to the kind of mental activity required by the various forms of item. Each item must consistently describe the behavior which is contemplated by each instructional objective. True-false items would ordinarily be regarded, for example, as inadequate for de-

termining the extent to which pupils have developed ability to organize information. Of the two basic forms of test item, recognition and recall, one form may be more effective than the other in controlling and directing desired types of performance. Careful study should be made of some standard textbook dealing with the mechanics of test construction in order to become familiar with the characteristics of various types of test item.

The manner in which a test is administered is indirectly related to the quantification of achievement. In the case of certain instructional objectives, appropriate emphasis upon rate and accuracy may be crucial. Tests may be given on a *power* or a *speed* basis. In the former case, time may be so unpredictable a factor in achievement as to make it undesirable to specify time limits. This condition exists when test items vary in difficulty or are purposely presented in an ascending order of difficulty. Time as a factor in achievement cannot be dismissed as totally irrelevant. If desired performance cannot be demonstrated within liberal time limits, the testing situation may be presumed to be too difficult. Certain behavior cannot be displayed adequately if insufficient time is allowed. With unrestricted time, it may be expected that the individual will complete his task if he is able to do so at all.

In some speed tests, the level of difficulty is uniform throughout; and the criterion of performance is the amount of the task accomplished within time limits. Usually the task is so designed that no individual completes it. Scores indicate the amount of work which each individual can complete within the allotted time.

Another type of speed test is one in which the time required for the individual to accomplish a certain task is measured. This type is called a "work-limit" test. A pupil's proficiency might be measured on the basis of the time required to typewrite one page. On the basis of the time usually required for an individual who has practiced for three months, criterion performance for satisfying a course requirement might be established. Ordinarily, raw scores would be in terms of time, with the shortest length of time representing the highest order of merit.

Tests may be designed with reference to the "breadth" of an individual's ability. For some purpose, it may be necessary to evaluate the range of his knowledge. On the other hand, it may be desired to determine the "depth" of his knowledge within a specified small area.

A test is always concerned with ability to do something. Quantification is usually achieved by some *indirect* process, since ability cannot be measured by a *fundamental* process. Responses may be

concerned with display of ability in three directions: (1) how much or how many, (2) how well, and (3) how rapidly. Such are the basic dimensions of ability. Any ability involved in performance related to an instructional objective may be explored from these three points of view.

Summary. The purpose of this chapter has been to outline the principles of deriving numerical values found in the various means of gathering data—the instruments of evaluation. Effort has been made to discover quantitative expression for the data that we are to collect. The chapter deals with data-gathering devices that may already be available or those that must be constructed prior to undertaking a quantitative study. These data-gathering devices include observational methods, the critical behavior technique, personal interview, the questionnaire, rating methods, inventories, and tests. Accuracy of our measurement will depend upon the extent to which we are able to quantify various phenomena to be studied. Adequacy of our instruments will depend upon the care with which they are constructed and refined and upon the needs and purposes of a particular investigation.

CHAPTER IV

Criteria of Measuring Instruments[1]

During the initial stages of planning a study, the investigator's typical procedure is to survey the results of other research which has a bearing upon his problem. Frequently, he discovers before he has unnecessarily given time and effort to the construction of an instrument, that an appropriate one already exists. Instruments published by commercial houses are available in quantity.[2] Many valuable instruments however have been published and described only in psychological or educational journals. Frequently these are worthy of consideration even though they may have been only partially validated and standardized.

In the study of many problems the investigator often will prefer to use already available instruments. Frequently such instruments have been standardized on the basis of representative cases and their validity and reliability empirically demonstrated. In many areas, instruments have been expertly constructed and standardized, particularly in relation to the measurement of such qualities as ability, achievement, aptitudes, personality, attitudes, and interests. Some of these instruments, even though they may not be entirely appropriate to a particular purpose are still superior to any that may be constructed with limited resources and facilities.

Many measuring instruments, although valuable for general uses, may not be appropriate for a particular purpose or problem. In

[1] No effort will be made to apply these criteria to all instruments of research—not even all of those discussed in the preceding chapter. The chapter deals with criteria that are generally applicable to any instrument regardless of type.

[2] Oscar K. Buros (ed), *The Third Mental Measurement Yearbook* (New Brunswick, N. J., Rutgers University Press, 1949).

experimental situations in which one wishes to determine the relative effectiveness of two or more methods of teaching, for example, the investigator may find it necessary to devise his own instruments for measuring pupil gain peculiar to this situation. New tests appropriate to the objectives sought under various methods may be essential to the investigation. In all such cases the investigator must either adapt existing instruments to his purposes or construct new ones. The needs may include a wide variety of tests, observational techniques, questionnaires, rating scales, and scoring cards.

It is always desirable to determine what a particular instrument will do toward accomplishing the aims that are sought. Will the particular instrument function in the collection of data needed in solving the problem defined? Can the abilities, traits, skills, information, or attitudes that are of interest to the investigator be measured adequately by the instruments considered? What criteria should be designated as evidence of successful or unsuccessful performance? What values are to be assigned to data that may be derived from use of given tools of research?

Whether the problem is one of selecting available instruments or of constructing new ones for a particular purpose, certain commonly accepted standards should be observed. Among the more important of these are *objectivity*, *validity*, *reliability*, and *discrimination.*

OBJECTIVITY

Objectivity in testing situations has reference to the fact that the same person may be expected to assign to an individual the same score on two or more occasions, or that different persons may be expected to agree on the score assigned to an individual. If the testing instrument has been so designed that it is capable of being scored objectively, it is possible to make fairly accurate comparisons of results of tests that have been administered by different individuals in different localities.

In connection with testing instruments, objectivity in scoring may be contingent upon what is referred to as objectivity in the meaning of the items themselves. Is a particular item free of ambiguity, or can it be variously interpreted? If more than one relevant interpretation is possible, the item must be considered faulty. The use of ambigious items seriously affects the extent to which a test may be scored objectively. Objectivity in meaning is dependent upon careful formulation of items and their editing and revision. Actual tryout of an instrument on appropriate populations is

needed in order to determine whether it has been constructed to attain the highest objectivity possible in scoring and meaning.

Objectivity in questionnaires refers to the extent to which respondents agree upon what facts are sought. In cases in which questions of fact are personal, responses necessarily are unique for each person concerned, and agreement among respondents cannot be expected. In items requiring expression of opinion or judgment as determined by rankings or ratings, however, the extent of agreement of individual responses should be checked. Smith says:

> The court not only wants to know the extent the witness can agree with himself in his several repetitions of his story; it also wants to know the extent of agreement of the different witnesses who testify. The scientist demands that the experiment be repeated, both by the original investigator and others. If the original investigator gets the same results a second or third time that is important. It is not sufficient, however; the investigator may be making the same mistakes each time. Other investigators may find these mistakes. If they find none, and if their results agree with those of the original investigator, generalizations begin to grow from hypothesis to theory or law. The greater the agreement among different investigators, the greater the confidence in the validity of the data and in the conclusions growing out of them.[1]

It is also important to determine extent of agreement of group with group. In the case of composite responses, individual variations may cancel each other; variable errors in one direction may cancel variable errors in the opposite direction. It may be expected that average interagreement of group with group will be closer than the interagreement of individual with individual. In questionnaire studies in which individuals or groups agree closely, it may be assumed that respondents are using similar standards in forming their opinions. Consequently, conclusions that they reach may be assumed to be derived objectively.

A frequently used method of discovering extent of interagreement of individuals is to analyze questionnaire items that require ranking. By using accepted average intercorrelation formulas it is possible to calculate average agreement of individual with individual. By using such statistical formulas as a basis for analysis, widely variable results are sometimes obtained. Individuals frequently do not agree closely with other individuals in their rankings. In many studies, however, in which low interagreement has been found the respondents were not experts. When all members

[1] F. F. Smith, *Criteria for Estimating the Validity of Questionnaire Data* (University of California, Berkeley; Thesis, 1932).

of a group are assembled and given instructions regarding the distinctive features in a situation, reasonably close interagreement usually results.

In rating scales objectivity is partially ensured by defining and describing the traits, qualities, or skills to be rated in such a way that each rater knows what they include and exclude. Raters also profit by rating under supervision; variability in opinion is thereby reduced. Similarly in observational methods, objectivity is increased by training observers in methods of observation and by familiarizing them with the differentiating qualities or distinctive features in an observational situation.

VALIDITY

An instrument is regarded as valid if it serves the purposes for which it is designed. This concept, despite its brevity and apparent clarity, means little until we have analyzed the conditions that applied when the instrument was validated. Some of the conditions for validation are concerned with certain logical methods of planning; others require refined statistical analyses of variable factors. Two aspects of validity to be considered are *logical* validity and *empirical* validity.

Logical validity. Logical validity is obtained when an investigator defines and describes the abilities, traits, concepts, or skills that he expects to be measured by an instrument of research, analyzes them to identify the elements needed in a measuring instrument, and designs the instrument with the demands of the situation as his criteria. If, for example, the investigator is confronted with the problem of building a test for measuring outcomes of a high school course in American history, it will be necessary to inquire what have been the specific instructional objectives of such a course, what specific materials have been used during instruction and study, and what particular activities have been directed. If instruction in the course has stressed acquisition of significant information in American history and certain cause-and-effect relationships, tests for measuring the results of instruction should be so designed that these two outcomes are measured.

Empirical validity. If we know that an instrument correlates closely with a selected criterion, we are at once confronted with the question of what the criterion itself measures. Suppose, for example, we correlate scores of a "problem-solving test" with another test of the same type and discover that the results are closely correlated. We must also undertake to determine what the criterion actually

measures, because as research workers, we are obligated to establish the worth of the criterion as well as that of the predictive instruments under consideration.

Empirical validity may be determined by two general methods: (1) the method of internal consistency widely used in testing situations and (2) the method of outside criteria.

Internal consistency in testing situations. The method of internal consistency in testing situations consists in showing the relationship between an individual's performance on a total test and his success on the various items that constitute the test. For example, after the administration of a test the investigator may calculate the relative standing of different members of his group, thus designating certain subjects as belonging to upper and lower positions of a frequency distribution. He may then determine the extent to which subjects located in the upper and lower positions of the distribution pass successfully individual items of the test. The assumption is that if the test is valid, there should be a larger percentage of subjects in the upper end of the distribution passing a certain item than those in the lower end.

A commonly used item analysis technique is that by Davis [1] in his "item analysis data" and employing for upper and lower groups 27 per cent of the entire group, thereby discarding at these points the middle 46 per cent. This method has an advantage over others (successive quarters, upper-lower, etc.) in that tables permit the items to be ranked in order of discriminating ability. Stanley [2] has devised a procedure for obtaining item discrimination indices which requires no computations other than simple addition and subtraction and which can be handled in its entirety by a reasonably conscientious assistant. For a typical 100-question teacher-made test the correlation between indices secured by means of this simplified method and by an "exact" technique was found to be .96, higher than the correlation between two different "exact" methods.

The fact must be taken into account that low representation of a certain kind of item in a test predisposes to lower discrimination indices for that type of item. Consequently, care should be exercised not to eliminate from the final edition of a test all items that do not satisfy the criterion of discrimination if a certain percentage of items of such a type is required in a test outline. For example,

[1] F. B. Davis, *Item Analysis Data: Their Computation, Interpretation and Use in Test Construction,* Cambridge. Harvard Graduate School of Education, 1946.

[2] Julian Stanley, *Short-cut Method for Estimating the Reliability Coefficient of a Test.* (As yet unpublished.)

consider a test consisting of 90 arithmetic computation exercises (involving three-column addition) and 10 problems involving arithmetical reasoning. If we adhere to the results of an inflexible analysis of items, reasoning problems might fare badly if they were not highly correlated with problems of computation. They should be included, however, in a final edition of the test if they have been listed as one of the objectives of a course in arithmetic.

In a few extreme cases some of the items may be disposed of with equal success by pupils in each of the various positions in a distribution, or there may be only slight differences among individuals in such positions with respect to supplying responses that are correct. It is also possible that some items will be passed by 100 per cent of the group, whereas other items may be failed by all members. In addition to the much discussed issue of whether to try to edit such 100 per cent and zero per cent items to make them easier or more difficult, there is the consideration that "padding" items of these types is likely to contribute nothing to the reliability (and nothing therefore to validity) of the test and may actually detract from its value. From the standpoint of results obtained per unit of time used, a test is superior without indiscriminatory items unless it is a mastery test under consideration, in which case they may be useful. Whether such items contribute anything to validity can be determined only by empirical data for the particular test under consideration.

Another procedure is to compare performance of individuals on each item with a criterion of what the test is designed to measure. When this method is used performance of persons scoring high or low on the criterion is compared with respect to their performance upon each item of the test. Biddle [1] validated his inventory by item analysis with teachers' judgment as the criterion and the extent to which the inventory differentiated between degrees of adjustment.

Correlation with an outside criterion. In correlating scores obtained from the use of an instrument with an *outside criterion* the investigator makes a direct attack upon the validity of his instrument. He is concerned with the degree of relationship between his instrument and some criterion of known or assumed validity. If a test, for example, correlates closely with a criterion it may be assumed that it measures in part the same qualities. On the other hand, if the degree of relationship is low, there is likelihood that it measures different qualities. Suppose the investigator wishes to construct a test to predict success in plane geometry. After he has

[1] Richard A. Biddle, "The construction of a personality inventory," *Journal of Educational Research*, 1948, *41:* 366–378.

been guided by certain logical considerations in constructing his test, he may administer it to students who, up to the time of the test, have not studied plane geometry. In such a case he will in all likelihood correlate scores made on his *Aptitude Test in Geometry* with course marks or with scores made on other standardized achievement tests adminstered as part of an end-of-the-course examination. If he obtains a satisfactory coefficient of correlation between scores on his *Aptitude Test in Geometry* and accomplishment as measured by course marks or standardized achievement tests he may conclude that his test is valid.

In the example just cited the investigator might use one or both of two criteria-course marks and a standardized achievement test upon the assumption that each of these two criteria is a measure of success in plane geometry. Obviously, he might have chosen other criteria, such as marks in previous mathematics courses, intelligence test scores, and teachers' estimates of efficiency.

COMMONLY USED "OUTSIDE" CRITERIA

Among the criteria frequently[1] used in validating measuring instruments are the following: (1) the outcome of an activity—such as failure and success in school or in vocational situations, (2) another measurement possessing known or assumed validity, (3) associates' ratings, (4) self-ratings, (5) factors isolated by factor-analysis techniques, and (6) responses of selected groups such as inmates in an institution or members of vocational groups. These several types of criteria will be briefly reviewed.

Outcome of an activity such as failure or success in vocational situations. If we wish to know whether a test will predict salesmanship ability, we may administer it to new employees of an insurance company and after a designated period of time correlate the test scores with the amount of insurance sold. The criterion would be the *amount of insurance* sold. Other examples might include extent of relationship between a *prognostic test of teaching efficiency* and degree of success on the job as measured by criteria such as ratings by the principal or the character and amount of pupil gain.

Use of similar measurement of known or assumed validity. Frequently authors of group tests of intelligence report coefficients of correlation between their tests and the *Stanford Revision of the Binet,* which has become a standard criterion for tests that stress

[1] E. H. Hsu, "A note on and some suggested methods for the determination of the validity coefficient," *The Journal of Educational Psychology,* 1948, *37:* 305–309.

measurement of abstract ability. If a particular group test of intelligence correlates closely with this test, it is assumed that the two tests measure similar abilities.

A similar procedure is often used in validating personality and adjustment inventories. The author of an inventory may correlate results of his tentatively constructed instrument with those obtained by other inventories bearing the same name. Some investigators find very little relationship between results obtained on one inventory and those on another indicating that particular traits are not easily defined or that authors disagree widely with respect to the characteristics being measured.

Ratings by associates. Ratings not only constitute instruments for collecting data but are also frequently used as criteria for validating varied types of data-gathering devices. Many authors of personality and adjustment inventories have validated their instruments by correlating the scores on their inventories with ratings of associates. The usual procedure is to request a number of persons who know particular individuals well to rate them on the same traits measured by the inventory. Ratings of the several associates are averaged and the averages correlated with scores made by the individuals upon the inventories that have been administered.

The results are usually expressed as validity coefficients. These coefficients are frequently not so high as might be desired, yet in view of the intangible nature of traits of personality and adjustment they are regarded as significant. Under favorable conditions —such as when, for example, raters are given training in the meaning of the traits to be rated—the coefficients obtained approach those found in validating *tests of ability*. Rating criteria are regarded valid to the extent to which ratings by different groups correlate closely with each other.

Self-ratings. Ratings by another person are usually regarded as more trustworthy than self-ratings. Nevertheless, self-ratings often provide a useful criterion. In reality, a person who takes a personality or adjustment inventory is rating himself by answering questions provided in such inventories. Yet, in taking such an inventory, the individual may be unaware of the traits being measured. After he has taken an inventory he may be given a list of the traits measured by it and asked to rate himself. When he is rating himself on traits that have not been explained to him, relationships between inventory results and self-rating may not be close. In rating himself upon traits the nature of which is fully disclosed to him, the individual frequently reveals an understanding of himself superior to any that may be obtained inferentially from the results

of an inventory. Self-rating often results in a more accurate revelation of traits than can be obtained through the indirect approach of an inventory.

Self-appraisal methods are coming to be highly regarded in many places. Pupils, for example, may appraise themselves on their work as well as enlist co-operation of their classmates in appraisals. Self-appraisals tend to create in such instances an attitude of co-operation and a sense of fair play, particularly when a teacher encourages self-appraisal by suggesting use of progress charts, permits pupils to correct their own errors in tests and examinations, and in other ways stimulates individual effort and initiative. The belief is common that a person is likely to be unable to make an objective, unbiased appraisal of his own qualities. Self-rating, however, in many instances may coincide closely with ratings by one's associates.

Factors isolated by factor-analysis techniques. Application of factor-analysis techniques to results of test performance has provided a means of making a direct attack upon the problem of empirical validity. Authors of tests are becoming increasingly concerned about the extent to which various parts of their tests (as well as the tests as a whole) isolate and measure, as such, relatively independent factors. Factor-analysis techniques make it possible for an author, for example, to determine whether his test is loaded with a number factor, a verbal factor, or with other factors. After a particular ability is believed to have been isolated and identified an important consideration is whether it is actually a relatively independent pure ability or trait instead of a blending of several abilities or traits. If it can be shown that the different parts of a test, or a test as a whole, measure traits that are distinctive or unique, we have a basis for a specific criterion.

Factor-analysis techniques are now widely applied to varied kinds of instruments, such as tests of motor skill, intelligence tests, special aptitude tests, and personality inventories. As a result of their use, authors of tests are able to provide those who use them with more exact information concerning their component elements than formerly.

Whether an investigator will wish to use a test that has been "factored" or one that tends to represent the conventional method of testing will depend upon his particular purpose. There may be justification, for example, for using a "general intelligence" test of the conventional type if the aim is to obtain a score that represents a survey of proficiency in varied types of activity. Although many conventional tests may consist of different parts and thus give the

impression that different abilities are measured, the different parts may actually overlap widely, thus making possible considerable intercorrelation coefficients. However, for the purpose of providing a score which is predictive of efficiency in some types of curriculum material such overlapping may not be a limitation.

Performance of a selected population as criterion. Frequently we find in psychological testing effort to validate a test by comparing scores of selected groups that may be expected to vary widely in their performance. For example, in the case of a mechanical aptitude test, the author may compare the averages of a group that are already successfully employed in mechanical industries with those of a group of persons who are just beginning employment in that field. He may also use the prevalent characteristics of experienced employees, with respect to their efficiency as shown by supervisors' ratings, in order to compare those who are rated *superior* with those who are rated *average* or *normal* in their performance. In both instances the performance of the samples selected affords a basis for determining empirically the validity that may be expected in the case of unselected samples.

This method may be applied in questionnaire construction where the purpose is to validate opinions expressed in response to various items. Private school pupils may be expected to answer certain questions in one way and public school pupils in another way. We may also expect church influence to be reflected in certain attitudes that are different from those of nonchurch groups.

VALIDATION METHODS ILLUSTRATED

The validation problem is essentially one of making a prediction on the basis of certain qualities or factors that are known or assumed to possess predictive values. Inasmuch as validation in psychological situations is concerned with the problem of making a prediction on the basis of present or past characteristics, the relationship between the traits or characteristics of the individual and his later adjustment or accomplishment must first be determined. The problem is one of determining the factors that are related to success in an activity so that these relationships may be used to forecast a particular individual's chances for success prior to his engaging in it. For such prediction to be possible the members of the group used for validation purposes must differ among themselves in their ability to perform in the activity concerned. It is also essential that a satisfactory criterion of success or failure in that activity be established. The criterion measure must be reliable and

adequate if predictions are to be accurate. A satisfactory criterion is of primary importance.

After an accepted criterion has been established it is necessary to identify, select, and measure the prediction factors that are associated with individual differences in the performance of the activity. We may then determine the degree of relationship between the predictive variables and the criterion measure. The closeness of agreement between the predictive variables and the criterion used as a measure of success is indicative of the degree of accuracy or the extent of the validity obtained.

Prediction factors in psychological situations may be classified into general categories: (1) personal factors and (2) situational factors. Personal factors include those physiological or psychological characteristics that pertain to a person, whether they be predominantly environmental or hereditary. Situational factors include those that are relatively independent of the individual, such as his educational or vocational background, his economic status, and the various community, national, and international influences that are outside his immediate control. Because of the number and complexity of these personal and situational factors the prediction problem is complex.

Prediction factors in a particular study will be as varied as the instruments that have been devised for measuring different aspects of an individual's development, including his physical status, intellectual development, interests, attitudes, and certain biographical information obtained through schedules. The greater the degree to which we are able to quantify the varied aspects of an individual's development the greater the accuracy of prediction.

Predictive indexes. Predictive indexes frequently include personal history data, educational records, general intelligence tests, achievement tests, special aptitude tests, personality and interest inventories, and combinations of factors.

In Table 2 are shown a number of scholastic aptitude tests which have been correlated with marks earned by engineering students at various stages of their training. Although these results suggest that several predictive indexes used singly are significant, no single index is sufficient for use as a guide for prediction.

Combinations of predictive factors. In establishing validity, predictions usually are not made on the basis of a single index but on a combination of two or more indexes. Statistical determination of the effectiveness of various combinations of indexes necessitates use of the method of multiple correlation and the computation of the multiple-regression equation. When this equation is used, each

TABLE 2. Relationship of Scholastic Aptitude Test Scores to Success in Engineering Training (adapted from Stuit and others: 1949)[1]

Study	*Predictive Index*	*Criterion*	*N*	*r*
1	ACE Psychological Examination	First-year honor-point ratio	154	.21
2	Thorndike Intelligence Examination for High School Graduates	Total four-year grade-point average		.43
3	University of Washington Intelligence Examination	First-year grades	193	.37
4	ACE Psychological Examination Q-score	First-year average	52	.41
5	ACE Psychological Examination L-score	..*do* *......	52	.28
6	Penn State College Psychological Examination:			
	Number Completion	First-year av.	132	.25
	English Usage	..*do*.......	132	.40
	Scientific Information	..*do*.. ..	132	.35
	Arithmetic Problems	..*do*.......	132	.39
7	Yale Scholastic Aptitude Test Battery:			
	Verbal Comprehension	First-year av.	120	.31
	Artificial Language	..*do*.......	643	.36
	Quantitative Reasoning	..*do*.......	643	.50
	Spatial Visualizing	..*do*.......	643	.38
	Mathematical Aptitude	..*do*.	643	.51
	Mechanical Ingenuity	..*do*.......	643	.31

* The abbreviation "do" has been substituted for ditto marks throughout the tables in this study.

variable is weighted in such a way as to yield the best prediction of the criterion. A procedure frequently used is to calculate first the prediction for each single variable with the criterion measure and then calculate the multiple-correlation coefficient (designated "R"), which shows the relationship between the best combination of predictive indexes and the criterion of success. The multiple-correlation coefficient, in special instances, may not be much larger than the best single index. However, as a result of errors of meas-

[1] D. B. Stuit and others, *Predicting Success in Professional Schools* (American Council on Education 1949, Washington, D. C., pp. 38–39).

urement and sampling, it is never smaller than the best single index.

Illustrations of the use of single indexes and combinations of predictive indexes are given in Tables 2 and 3, which provide a summary of research relating to prediction of success in engineering. Table 2 shows the effectiveness of the use of a single index, whereas Table 3 shows the effectiveness of using combinations of variable factors.

In Table 3 are shown the results of several studies when combinations of indexes are used in predicting the criterion of success in engineering. The results as summarized would seem to indicate that accuracy of prediction will not necessarily be increased by use of any combination of several variables. In some studies the multiple coefficients may be slightly less than zero order coefficients between achievement in engineering school and one variable. In other instances multiple-correlation coefficients may be of the same size, regardless of whether a few or several predictive indexes were used. It would seem from these and other data that the most effi-

TABLE 3. Multiple-Correlation Coefficients Showing Relationship among Various Combinations of Predictive Indexes and Success in Engineering Training (adapted from Stuit and others: 1949)[1]

Study	*Predictive Index*	*Criterion*	*N*	*R*
1	Columbia Research Bureau: Chemistry Test, Physics Test, Algebra Test, Plane Geometry Test	Four-year grade-point average		.58
	Thorndike Intelligence Examination for High School Graduates Columbia Research Bureau: Algebra and Geometry Tests Cox Mechanical Aptitude Tests: Models and Completion Minnesota Paper Form Board Test Minnesota Interest Analysis Blank	..do.......		.59
	Cox Mechanical Aptitude Tests: Models, Completion Explanation	..do.......	104	.46

[1] D. B. Stuit and others, *Predicting Success in Professional Schools* (American Council on Education 1949, Washington, D. C., pp. 38–39).

Study	Predictive Index	Criterion	N	R
	MacQuarrie Test for Mechanical Ability: Copying, Blocks, Location, Pursuit, Tapping Minnesota Paper Form Board Test Cox Mechanical Aptitude Tests: Models	..do.......	104	.46
	Cox Mechanical Aptitude Tests: Models	..do.......	104	.44
2	Iowa High School Content Examination Iowa Silent Reading Test Iowa Placement Examination Series: Mathematics Aptitude Test, English Training Test	First-semester, freshman-year, grade-point average	99	.77
3	Iowa Placement Examination Series: Mathematics Training Test Co-operative Intermediate Algebra Test ACE Psychological Examination, Thurstons Primary Mental Abilities Series, "V"-factor Rank in high school graduating class	First-year grade-point average	200	.76
4	ACE Psychological Examination Purdue Placement Test in English Iowa Placement Examination Series: Mathematics Training Test	First-semester, freshman-year, grade-point average	1,113	.72
5	University of Washington Intelligence Test Iowa Placement Examination Series: Mathematics Aptitude and Training Tests, Physics Aptitude and Training Tests High School average in English, Natural Science, Social Science, Mathematics	First-year average	193	.68

cient combination of predictive indexes for predicting success in engineering school includes (1) previous scholastic record (high school or college), (2) scholastic aptitude test scores, and (3) scores obtained on achievement tests, especially in the areas of mathematics, science, and English.

The predictive value of intelligence tests. Some of the best examples of validation are found in the area of ability testing. The general intelligence test, as the term implies, attempts to measure the learner's reaction to varying types of material so that the total score resulting from composite treatment of its various sections indicates the student's potential learning ability in a variety of learning situations.

In many instances, the aim is to select testing materials that reveal the learner's performance in different situations. Consequently, in order that the total score resulting from composite treatment of the several sections of a test may correlate significantly with "outside" criteria (such as demonstrated achievement, teacher's estimates, or other tests of known or assumed validity), performance on the different sections should not correlate closely. Low intercorrelations among different sections of a test suggest that different mental functions operate; high intercorrelations imply that similar mental functions are involved.

General intelligence tests have tended to stress the measurement of higher mental processes, emphasizing the learner's ability to react to verbal materials. Thus they tend to be heavily weighted with items requiring language responses as opposed to motor reactions. The belief that tests of general intelligence should be measures of ability to think by means of symbols has been generally accepted in constructing tests for use in secondary schools and colleges. And this emphasis upon abstractions and symbols has made such tests particularly useful in predicting general scholastic efficiency.

The more closely test materials coincide with the materials and mental processes used in learning situations in school and college, the greater their predictive value. Psychological tests that stress measurement of verbal ability correlate with achievement in verbal subject matter as highly as 0.50 or 0.60; those that minimize the verbal factor, on the contrary, may not correlate more than 0.30 or 0.35. These results apply, of course, to the total score of a test. In contrast, if predictions are desired in more specific learning situations, as for example, in foreign languages and mathematics, sections of the psychological examination that emphasize language and numbers will be even more highly predictive of performance in those subjects.

Measurements of special abilities include aptitude tests in fields such as music, art, and mechanics; and prognostic tests in subjects such as mathematics and foreign languages. Aptitude tests have tended to be more particularly concerned with vocational fields; prognostic tests, with school subjects. In each case, however, the objective has been to devise testing situations that will be indicative of the learner's efficiency in specific situations. Inasmuch as aptitude and prognostic tests are designed for specific purposes, and testing materials so selected as to be representative of types of performance sought, it is by no accident that they tend to have greater validity in specific situations than general psychological examinations, which usually measure superficially a wide range of abilities. Scores on some general psychological examinations, for example, correlate with achievement in geometry 0.50 or 0.60. A prognostic test in geometry on the other hand, may correlate with accomplishment in geometry as highly as 0.70 or 0.75. And in the case of certain jobs and trades, in which activities are reduced to specific skills, validity coefficients may be as high as 0.75 or 0.80.

The validation of personality and adjustment inventories. Criteria used in validating personality and adjustment inventories include intelligence tests, ratings by associates, school marks, and inventories designed to measure similar traits. A method frequently used in validating personality and adjustment inventories is that of correlation with ratings. Ratings provide a means of obtaining certain information that can be obtained in no other way. The ratings should yield quantitative scores; the raters should be conscientious and trained in the techniques of rating.

Another method of validation is one that considers the extremes of a distribution of scores and disregards the middle ranges. In two widely separated groups, two extremes are selected. The internal validity of an inventory is determined by the extent to which it makes a similar differentiation. Two groups are examined to determine whether a battery of tests will separate the groups with slight overlapping. An advantage of this method is that it affords a basis for evaluating each item of the inventory. The item is valid to the extent that it differentiates between the groups. One limitation of this method is that it is often difficult to obtain two extreme groups; another is that it fails to show how well an inventory discriminates in the middle ranges.

Ellis,[1] in a comprehensive analysis of research involving personality inventories, concludes that for every study of validity there

[1] Albert Ellis, "The validity of personality questionnaires," *Psychological Bulletin*, 1946, *43:* 385–440.

are approximately three or four involving reliability. Of the relatively few studies involving validity a majority have not been objective clinical validations; instead they have consisted of statistical checks as part of the particular method of test construction. For the most part they have consisted in little more than item analysis by the method of internal consistency.

When direct validation methods are considered (that is, those that consist of using outside criteria such as clinical diagnosis and ratings of behavior problems, psychiatric and clinical diagnosis, and ratings by friends and associates), Ellis finds that out of 162 studies examined 65 show positive, 26 questionable positive, and 71 negative results. On the basis of his analysis Ellis concludes:

> We may conclude therefore that judging from the validity studies on group administered personality questionnaires thus far reported in the literature, there is at best one chance in two that these tests will validly discriminate between groups of *adjusted* and *maladjusted* individuals, and there is very little indication that they can be safely used to diagnose individual cases or to give valid estimation of the personality traits of specific respondents. The older and more conventional of these inventories are hardly worth the paper they are printed on.

Ellis and Conrad [1] in a further review of personality inventories as applied in military practice are more optimistic. In military practice these inventories have been validated in two ways: (1) by use of a psychiatric criterion in which the scores of "normal" groups of enlisted men and officers are compared with the scores of others who have been diagnosed as neuropsychiatrically unfit; and (2) by use of a measure of performance comparison of inventory scores of those who were successful in training or combat with those who were unsuccessful. The authors conclude that although a cautious attitude regarding the results obtained should be encouraged, these inventories made a substantial contribution to the screening process in military practice and that the results should prove helpful in civilian situations.

The authors draw the following conclusions:

> 1) Personality questionnaires should be especially designed for the group to whom they are applied and should be validated against dependable external criteria. Criterion contamination should be guarded against; and criterion overlap, if it occurs, should be taken into account in evaluating the findings.
>
> 2) Special attention should be given to persuading or inducing respondents to answer the inventory items as truthfully as they can.

[1] Albert Ellis and Herbert Conrad, "The validity of personality inventories in military practice," *Psychological Bulletin*, 1948, *45:* 385–426.

3) Personality inventories may possibly be more effective when used with relatively uneducated and less intelligent groups than with groups that are more sophisticated.

4) The users of personality inventories should realize that only limited and specialized demands may be made on the inventory technique; and that broad and incisive personality diagnosis is still the specialty of the trained clinician employing subtler and more comprehensive techniques.

It may be expected that such inventories and questionnaires will be improved both in their administration and in the technique of requiring responses. One improvement in eliciting responses includes use of the *forced-choice technique,* which is so designed as to force the subject to select items from two or more different continua or categories in paired or triad form. Improvement is also noted in the assignment of the empirical weights based upon performance of contrasted groups.

Personnel directors and placement officers attach considerable significance to personal biographical data supplied by the applicant for a job. When such data are accurate and include significant items in one's career, they afford a valid basis for making certain predictions with respect to a person's educational or vocational fitness in particular situations. These personal histories have been shown upon analysis to have relatively low validity in "cross-validation" groups. Items relating to biographical data, however, have proved to be reasonably successful in predicting success of life insurance salesmen, in selecting personnel for the army, and in appraising candidates for admission to college or university.

When obtaining biographical data the investigator should exercise special care in formulating questions in order to prevent responses from being biased in favor of the respondent. He should also recognize inability of respondents to recall with accuracy significant aspects of their experience and record.

THE VALIDATION OF QUESTIONNAIRES AND RATING TECHNIQUES

Questionnaires. Widespread use of the questionnaire in collecting information in education has caused many potential respondents to react negatively to almost any questionnaire. They frequently feel their inadequacy in supplying the information required and feel that the time and effort needed to respond to a questionnaire place excessive demands upon their time. Frequently, investigators fail to create an attitude of co-operation

from the respondents during the initial stages of planning a questionnaire study. Often those who are capable and willing do not feel that they can spare the time or effort to reveal themselves in their true status. Some of the difficulty may be obviated by constructing recognition type items, thereby requiring only that the respondent check answers with which he agrees.

In some cases validity is likely to suffer as a result of inability of a respondent to supply precise and clearly formulated answers to types of item. The bias of the investigator may be unconsciously reflected in the items themselves thus making an objective response impossible for the respondent.

Inaccuracy is due, in many instances, to the sporadic and hurried manner with which questionnaire items are constructed. Questionnaires, when carefully planned in the light of the objectives of an investigation, the kinds of data needed, and with due consideration of the ability and willingness of potential respondents to supply data are capable of yielding reasonably accurate results. Questionnaires, like other instruments of research, not only should be carefully constructed and edited but should be subjected to empirical tryout on selected respondents prior to their use in an investigation. The validity of a questionnaire is increased in accordance with the amount of care, patience, and effort exercised in its construction.

Blankenship [1] believes the wording of questions is no less important than appropriate sampling and that the only way of making sure that questions are properly worded is to conduct, in advance, an experimental study on selected respondents. Blankenship suggests the following means of improving validity of questionnaire research:

> Thorough understanding of the specific problem must precede questionnaire construction, and various factors must be taken into account: the critical character of the first few questions; the avoidance of ambiguity; the use of words understood by the lowest class of respondent; the use of questions that are reasonable and concrete; the adapting of questions to the type of person interviewed; neutral phrasing of questions and avoidance of suggestion. With a questionnaire constructed according to these standards a pretest of 25–30 interviews will eliminate difficulties in phrasing. The interviewer will observe the results in terms of the criteria stated, and, if necessary, attempt variations in wording. For greater accuracy a sample study is needed.

[1] A. B. Blankenship, "Pre-testing a questionnaire on public opinion," *Sociometry*, 1940, *3:* 263–269.

Ratings. The rater's opinion, attitudes, and fund of general experience are all involved in the activity of rating. The synthesis of such a background into a rating may be regarded as a spontaneous act on the part of the rater. That is, his final conclusion may remain in a state of suspension not known even to himself, until the moment that he is required to make the rating.

As a result of these varying factors wide individual differences are exhibited when various raters rate the same characteristics, qualities, or traits. In rating individuals on certain traits or qualities these factors include the observer's general reputation, and the degree of acquaintanceship of the rater with him. Among the factors which influence ratings are the rater's past experience, and opportunity for observation and training.

Validity may be expected to vary with the type of scale, as for example, the *man-to-man scale*, the *graphic scale*, the *method of paired comparison*, and the *scoring card*. It is difficult to formulate precise rules indicating the relative validity of any of these scales. High validity values may be found under one set of conditions with a particular rating scale and low values under another.

In a rating scale, the more steps that involve fine distinctions, the more frequently are errors in judgment likely to result. Some authorities suggest that there should not be less than five nor more than seven steps on a scale for most satisfactory results. Too few steps would result in the measurement being too rough or approximate; too many would result in error because of the rater's inability to discriminate among fine shades of gradation.

Rating methods are never any better than the raters. We may refine our rating methods and administer them properly but unless the rater takes the job of rating seriously and conscientiously the results are of questionable value. A good rater is one who possesses the following characteristics: (1) he should have frequent opportunity to observe and experience the thing being rated, (2) he should be an expert in some area, (3) he must be prepared by preliminary instruction for the rating, he must be able to know what it is that is being rated, and (4) he must usually have some frame of reference in connection with his rating. Is he to consider "*average teachers*" to mean "*average*" for a particular size community or a larger size city school system? For example, does "*superior*" mean "above average" or what?

Most dependable results are obtained when there is a comparatively large number of raters. It is generally desirable to use a large number of raters if feasible and compute an average of the ratings

obtained. However, a large number of raters is not necessary in case each rater is especially well qualified, being for example unusually well informed in a field of investigation.

THE SIGNIFICANCE OF CRITERION MEASURES

A crucial problem in validating instruments of research is that of obtaining satisfactory criterion measures. The aim is to obtain a criterion measure that is statistically reliable and adequate for the purpose. Frequently, it is necessary that several criteria be used and that these criteria not resemble each other closely.

The value of a criterion measure depends upon the degree to which it meets the criteria of reliability, adequacy, and discrimination. Evidence of validity is manifested in the extent to which various measures of success are positively correlated. In some situations, however, it is possible that a number of criteria may be relatively independent of each other.

Mosier [1] has pointed out that a test may be its own criterion in which case the validity coefficient is simply the square of the reliability coefficient of the test. Often, however, we are concerned with an antecedent-consequent (predictor-criterion) relationship. When using one set of measurements to predict measurements in a different function—for instance, intelligence test score at the beginning of a school year and average marks at the end—we must recognize the *reliability* and *adequacy* of the criterion.

If the criterion is unreliable, the obtained correlation coefficient between predictor and criterion except for chance fluctuations will be lower than the intrinsic or "true" validity coefficient. Unreliability results from limited variability of criterion measurements and large errors of measurement. When either defect is present measurements fail to differentiate sufficiently among the persons measured.

Inadequacy of the criterion is a much more stubborn problem than unreliability. Unreliability can usually be reduced by careful editing of questions, item analysis, lengthening the scale, and other techniques. Jenkins [2] believes that in order to be optimally useful the criterion must be a summary measure indicating the sort of proficiency that the investigator is trying to predict. To argue the contention that the criterion itself must be "valid" seems almost to

[1] Chas. I. Mosier, "A critical examination of the concepts of face validity," *Education and Psychological Measurement*, 1947, *7:* 191–205.

[2] John G. Jenkins, "Validity for what?" *Journal of Consulting Psychology*, 1946, *10:* 93–98.

belabor the obvious. One would not consider a "speed of tapping test" suitable as a criterion of success in an English literature course, even though it might be highly reliable and easily obtained. That the criterion should be logically relevant and appropriate as well as "valid" from all technical points of views is a truism that is applicable thoughout our discussion of criteria.

Validation of intelligence tests is attended by pitfalls because no adequate criterion of "intelligence" exists. Teachers' marks, used widely, are contaminated with elements of promptness, conformity, industriousness, and socio-economic background; they are also usually unreliable. Subjective estimates, age in grade, scores on achievement batteries, scores on other intelligence tests, occupational level, and similarly proposed criteria all possess readily discernible defects. This problem sometimes has been partially resolved by narrowing the concept of test intelligence to refer to only relative success and failure in certain subject-matter areas and by changing test titles from "intelligence" to "scholastic aptitude," "academic aptitude," or just "psychological examination."

The criterion may be a single one, such as amount of insurance sold during the last twelve months, or it may be a composite of various factors, weighted according to their importance. Thus the kind of territory to which the insurance salesman has been assigned may be taken into account, together with his length of employment, his community activities, and the average size of policies sold. Although several procedures for determining optimal weights to be assigned the various components have been suggested, in practice, subjective judgment is usually employed. Additional considerations are necessitated when a dichotomous criterion (e.g. pass-fail, productive-unproductive, quit before end of first year vs. worked entire year) is used in lieu of a continuous one.

The extent to which the criterion measure discriminates should be considered. Designation of a number of degrees of success will usually have the effect of increasing the degrees of correlation between measures of success and the criterion measure. But a criterion broken into two or three levels of success is preferable to a multidimensional classification that is less reliable. The investigator should use as many categories as may be necessary to describe success in the particular situation so long as they can be reliably discriminated and meet the demands for accuracy in representation of data.

In evaluating the criterion of success it is necessary to determine the extent to which it is possible to evaluate the differences among the simpler predictions on which worth of the criterion of success

is based. Instruments constituting predictive variables should be reliable and valid. It also follows that the criteria as well as the instruments used as predictive variables should be sufficiently different from each other to permit discriminations. Inasmuch as it is difficult to know precisely what a criterion of success is we should select several criterion measures from those that may satisfy the needs of a particular situation.

Coefficients of validity may be misinterpreted as a result of errors of measurement. If it is desired to determine the extent to which a test predicts success in college or university, it is necessary to establish a criterion and devise an appropriate measuring instrument for predicting success. A high coefficient of correlation between a test and a criterion is often the result of a comparison with a criterion that may be highly fallible. A low coefficient of correlation may be due to a number of factors, including imperfections in the measuring instrument and defects in the criterion. For this reason it is sometimes desirable to determine the "index of forecasting" efficiency of a test, after correction for random errors of measurement in the criterion. Conrad and Martin [1] have suggested a formula by which such a correction may be calculated, and they have provided a table of values (the corrected index of forecasting efficiency) for various values (the correlation between test and criterion and the reliability of the criterion).

Some conditions of validity. The research worker should guard against accepting at face value coefficients of validity reported by authors of standardized instruments. The coefficients reported represent results obtained for a particular population under certain conditions. It is necessary that the author of an instrument report the empirical findings resulting from the standardization process. It should be recognized that different results might be obtained with different populations under different conditions.

Variability in validity coefficients may be expected according to the nature of the population including differences in maturity and in level of ability, experiential background, the nature of the criterion used for comparison, and other factors. A test designed for several grades, for example, may have higher validity for some grades than for others. Validity may also vary with the sex of the population, that for boys being in a specific instance higher than that for girls.

In considering validity we must never lose sight of the purposes for which certain instruments have been designed. We must raise

[1] H. Conrad and G. B. Martin, "An index of forecasting efficiency for the case of true criterion," *Journal of Experimental Education*, 1936, *4*: 231–244.

the question: "Validity for what?" For example, in almost any area of measurement—such as intelligence, special aptitude, interest inventories, or personality schedules—it is possible that a number of instruments are generally valid. In a sense each instrument constitutes its own definition; such definitions may be expected to vary from one instrument to another. Wesman states:

> How valid each test is depends on two considerations: whether you agree with the definition of intelligence as represented by the content of the test, and what you are trying to predict (and the ability to learn is certainly one good definition of intelligence), then that test is most valid which best predicts the given kind of learning. The validity of a test is always specific to a situation; a "generally" valid test is one that is satisfactorily valid in a large number of specific situations.
>
> The same considerations that determine the validity of intelligence tests also determine the validity of achievement tests: what specific ability we are trying to evaluate, and whether we agree that the test's content satisfactorily taps that ability. What is the validity of a spelling test? We need to define spelling, just as we need to define intelligence, before we can judge validity. What kind of spelling ability are we interested in? Is it the ability to single out incorrectly spelled words, as a proofreader or editor need do? If so, our test should be one that provides direct evidence of that skill. Or is it the ability to recall correct spelling of words for creative writing? In that case, the spelling test should tap that specific skill. It may be proposed, and quite legitimately, that the two kinds of spelling tests would correlate highly. Nevertheless, it has not been demonstrated, so far as the writer knows, that the correlation (even when corrected for attenuation) would be perfect. Since the skills are different, the difference should be taken into account when selecting the test to be used in a specific situation.[1]

Inasmuch as there can be no all-purpose validity the research worker should check the validity of an instrument that he is considering for his particular population in accordance with the specific conditions that affect the investigation. In this way he can establish his own criteria in accordance with the particular purpose of his study and thereby establish a coefficient that will be appropriate for the situation.

RELIABILITY

Reliability in testing situations. The fundamental concept of reliability resides in the assumption of consistency or stability in scores when there are repetitions of measurement with an instru-

[1] A. G. Wesman, "Needed: more understanding of present tests in improving educational research," *American Education Research Bulletin*, 1948, pp. 63–68.

ment. Does the subject in a testing situation, for example, earn approximately the same scores on two administrations of a test?

The reliability concept in educational situations differs in certain fundamental respects from that used in physical measurements. Cronbach [1] has called attention to the fact that the physicist assumes that repeated observations are independent in the sense that uncorrelated errors are at a minimum. The situation in physics remains constant during repeated observations and consequently possibility for error is much less in evidence than in the case of the educator, who must reckon with variations in time for administering a test, variations among the individuals studied, and variations in the task.

Thorndike [2] has proposed a number of questions regarding assumptions underlying the concept of repeated measurements. Among other things he discusses the question of the extent to which we wish to generalize with respect to the consistency of a test. Each inference that one makes with reference to the use of test scores influences our concept of reliability. The purpose for which test scores are to be used constitutes a practical guide for our judgment of the adequacy of their reliability.

Three concepts of reliability. The concept of reliability includes as "error variance" fluctuations in the performance of observers that may be attributed to variations in (1) period of time, (2) situations, and (3) observers. These variations in performance of individuals that may be attributed to fluctuations in performance during a *period of time,* a *sample of test tasks,* or *both* are expressed by three coefficients: (1) coefficients of equivalence; (2) coefficient of stability; and (3) coefficient of equivalence and stability.[3]

[1] L. J. Cronbach, "Test reliability; its meaning and determination," *Psychometrika,* 1947, *12:* 1–16.

[2] R. L. Thorndike, "Logical dilemmas in the estimation of reliability," Nat'l Proj. in Educ. Measmt., Am. Council on Educ. (Series No. 28, 11, 21–30).

[3] For assumptions underlying these concepts and technical details of computation examine the following:

L. J. Cronbach, "Test reliability: its meaning and determination," *Psychometrika,* 1947, *12:* 1–16.

D. C. Adkins and H. A. Toops, "Simplified formulas for item selection and construction," *Psychometrika,* 1937, *2:* 165–171.

L. Guttman, "A basis for analyzing test-retest reliability," *Psychometrika,* 1945, *10:* 255–282.

L. Guttman, "The test-retest reliability of qualitative data," *Psychometrika,* 1946, *9:* 81–95.

G. F. Kuder and M. W. Richardson, "The theory of the estimation of test reliability," *Psychometrika,* 1937, *2:* 151–160.

J. Loevinger, "A systematic approach to the construction and evaluation of tests of ability," *Psychological Monographs,* 1947, *61:* 1–49.

Coefficient of equivalence. The coefficient of equivalence shows the extent to which scores on two forms of the same test fluctuate when administered at one sitting. It shows how the individual's performance fluctuates when he is measured on two different samples of the same behavior. If the individual's accomplishment on one form of a test is similar to his performance on another, the test is reliable.

When a test is of such a nature that it is possible for a person to recall or recognize some items from the first to the second testing as in the test-retest method or when there is possibility of practice effect resulting in carry over from one testing to another, it is frequently desirable to arrange two equivalent forms that consist of different items selected as samples of the same abilities.

The coefficient of equivalence is frequently calculated by using what is known as the split-half method, in which the individual's success on odd-numbered items of a test is correlated with his success on the even-numbered items of the same test. This procedure has the advantage of ruling out possibilities of practice effect, fatigue, and other factors, and also of sidestepping the need for administering a test more than once. Inasmuch as one half of a test is correlated with the other half it is necessary when using this procedure to determine by the Spearman-Brown formula the reliability of the entire test. This formula makes it possible under certain assumptions to estimate the coefficient of reliability when the test is lengthened or shortened. The split-half method, however, may provide an inaccurate picture unless the two half tests are as equivalent as the two forms of the same test would be. The means and standard deviations of the two halves should be approximately equal. In estimating the correlation between scores on the two halves the method of maximum likelihood gives the best estimate. The halves should also be comparable with respect to content and difficulty of test material.

A variation of the split-half method has been recommended by Cronbach [1] who suggests wherever possible the use of what he calls the "parallel-split" method. When using this procedure the investigator makes no assumptions regarding the equivalence of odd- and even-numbered items but empirically determines the comparability of the two samples of behavior. A number of test papers are examined to determine the number of persons passing each item. The items are then classified into two groups in such a way that the two halves are approximately equal in content and difficulty.

[1] L. J. Cronbach, "Test reliability: its meaning and determination," *Psychometrika*, 1947, *12:* 1–16.

Another group of papers is then scored on the two half-tests and the appropriate formulas are applied.

The Kuder-Richardson [1] formula and its variants provide dependable estimates of test reliability when the parallel-split procedure has been arranged. The split-half method, on the other hand, either overestimates or underestimates the degree of reliability, depending upon the extent of comparability of the two forms of a test. Neither the split-half nor the parallel split procedure provides accurate estimates unless not more than a single common factor accounts for the inter-item correlations. If a test measures a number of different abilities the coefficients are likely to be too low. The split-half and the parallel-split methods are not applicable in the case of speed tests. Jackson [2] has developed a measure of the sensitivity of a test. Hoyt [3] has applied the analysis of variance to the determination of test reliability.

Coefficient of stability. As previously stated the basic concept of reliability is consistency of performance in repeated measurements. The coefficient of stability provides an estimate of the degree to which an individual's score will vary in the case of identical sets of test items during a period of time. These estimates of reliability tend to vary inversely with time intervals. For that reason Guttman [4] in his treatment of the problem takes such time intervals into account. He also derives estimates for the lower limit of this reliability coefficient for performance on a single trial.

Coefficient of equivalence and stability. The coefficient of equivalence and stability, as the term implies, shows the extent to which an individual is consistent in his performance on two comparable forms of a test over a period of time. Reliability is estimated by what is called the "delayed parallel-test" method. The coefficient reflects both the fluctuations in performance of the individual and his choice of specific items of the test. Two forms, comparable in difficulty and content, are administered to the same persons on two different occasions. By correlating the two sets of scores we arrive at a coefficient.

[1] G. F. Kuder and M. W. Richardson, "The theory of the estimation of test reliability," *Psychometrika,* 1937, *2:* 151–160. Kuder and Richardson have proposed the method of "rational equivalence," which is basically a measure of internal consistency in items. The formula takes into account the intercorrelation of individual test items.

[2] Robert Jackson, *et al.* "Studies on the reliability of tests," Toronto, Department of Educational Research (Univ. Toronto, 1941).

[3] C. Hoyt, "Test reliability estimated by analysis and variance," *Psychometrika,* 1941, 6:267–287.

[4] L. Guttman, "A basis for analyzing test-retest reliability," *Psychometrika,* 1945, *9:* 255–282. See also, "The test-retest reliability of qualitative data," *Psychometrika,* 1946, *10:* 81–95.

TABLE 4. Summary of Methods Used in Determining Reliability [1]

Name and Method of Determin.	*Assumptions*	*Error when assumptions are violated*	*Quests answered by coeff.*
1. Coeff. of equival. split-half	Halves must be equiv.	Coeff. for spd. tests falsely high. For other tests coef too low if halves not equiv.	How precise does test measure?
a) Kuder-Richardson	Test must measure a single factor	Coeff. for spd. test falsely high, for others falsely low if items measure many factors	How adequately does it sample all items that might be included?
b) Immediate parallel test	Tests must be equiv.	Coeff. shows degree of equiv. rather than accuracy of either test.	None
2. Coeff. of stability retest after interval	No opportunity for increasing ability by practice during interval.	Pr. on function decreases coeff.	How stable is meas. with test?
3. Coeff. of stability and equiv. parallel test after interval	Tests must be equiv. no opportunity for practice	Coeff. underestimates accuracy of test if assumptions violated.	How would samples of behavior at one time correspond to results from similar sample at another time?

Applicability of the three methods. Applicability of the three methods depends upon the needs of the testing situation. In a speed test the coefficient of equivalence should be used only when parallel test forms are administered immediately. The Kuder-Richardson and "split-half" methods should not be used in speed tests.

[1] Adapted from L. J. Cronbach, *Essentials of Psychological Testing* (Harpers, New York, 1949).

The coefficient of stability is applicable in cases in which one considers fluctuations from day to day as sources of error. In computing reliability of intelligence and aptitude tests that require one's immediate reactions to problem situations and are relatively unaffected by environmental influences, either the *coefficient of equivalence and stability* or the *coefficient of stability* is appropriate. Cronbach recommends that in cases in which fluctuations in performance are attributable to "real" variables instead of error the coefficient of equivalence should be used.

Factors affecting reliability in testing situations. The reliability of a test is dependent upon the reliabilities of its various items. The reliability of each item depends upon the skill and care exercised in its preparation. There is no adequate substitute for skill in writing and editing items that make up a test. Specific aspects of item construction are discussed at length by Adkins.[1]

After items have been carefully composed according to specifications outlined in advance, they should be tried out on a sample of subjects as similar as possible to those for whom the test is intended ultimately. Arbitrarily, we might stipulate that twice as many items as are to be retained in the final form be used in the preliminary edition and that the trial sample include several hundred subjects.

The revised test will contain re-edited items of the proper difficulty level (a complex consideration but frequently centering around 50 per cent when corrected for "chance") and that discriminate sufficiently well between high ability and low ability members of the group. In practice, we are usually concerned with the question of whether a given item is answered correctly by a significantly larger number of persons who score high on the entire test than by those who score low. If we can succeed in causing wide variability among total scores of the examinees, that is, have some go very low and others very high, and if each subject's score is typical of what he would have done on another similar occasion, the test will likely possess high reliability.

In addition to care in construction, editing, and revising items, and attention to their difficulty levels and discriminating power, there are a number of factors that should be recognized:

1) *Length of test.* Because each test represents a sample of behavior, it follows that as the sample becomes increasingly extensive the greater the opportunity for thorough comprehensive measurement and consequently the greater the degree of reliability. Many items provide increased opportunity to measure the subject's true

[1] D. C. Adkins, *et al.*, *Construction and Analysis of Achievement Tests* (Washington, D. C., U. S. Government Printing Office, 1947).

ability; opportunities for guessing also decrease in proportion to the number of responses made as the amount and range of material in a test increase.

Tests containing a small number of items invariably yield low reliability coefficients. A ten-item test will be less reliable than one containing twenty equally good items of the same kind if used with the same individuals. The investigator by means of the Spearman-Brown formula can estimate the reliability of the test when the length is increased or decreased. This formula applies, however, only when the longer and shorter tests are comparable with the one on which the estimate is based. It is assumed that the longer the time a test requires, within reasonable limits, the higher the reliability.

2) *Range of difficulty of test items.* Reliability of a test is unaffected by omitting items so difficult that no one in the group answers them correctly; neither is it affected by omitting items so simple that everyone in the group answers them easily.

Arranging test items in ascending order of difficulty by empirically determined methods increases reliability, particularly when this is done by uniformly successive steps of difficulty. "Bunching" difficult items together has a similar effect on reliability to that of decreasing the number of items.

3) *Ability level of the group.* To ensure satisfactory reliability a test must be appropriate for the group on which it is used. If a test is either very easy or very difficult for a group a skewed distribution results, and will be unreliable for members of the group as a whole. In some instances a test may be satisfactory for measuring individuals at the upper range of ability and unsatisfactory for those in the lower range. A test may possess high reliability for a group representing a wide range of differences and low reliability for a group representing a narrow range. For example, it may be highly reliable when scores for pupils in several grades are treated compositely and low in reliability when computed for pupils in a single grade where the range of scores is narrow.

There remain several other factors that influence reliability, as follows:

a) *Examiner's personality.* Directions for the test should be standardized thoroughly if it is to be administered to more than one individual, but even then the examiner's personality, temperament, and mannerisms may affect results.

b) *Format of the test.* Format of a test should be attractive. Crowding, illegibility, and confusing arrangement contribute to unreliability.

c) Time limits. Time limits, if used, should be determined by "try-out" rather than by rough estimates. If a test is designed to measure speed of reaction largely independent of accuracy, it will require timing different from what it would be if concerned primarily with accuracy. In the latter situation time limits serve the purpose of administrative convenience; prevention of restlessness, disorder, and wasting time of those who complete the test early. The amount of time is related to effective length of the test and may influence reliability. There is an optimal limit for a test when used in a given situation and for a stated purpose.

Reliability in rating techniques and questionnaires. Rating methods and questionnaires pose a somewhat different problem with respect to reliability than those encountered in testing situations but the principles are fundamentally the same.

Reliability of rating methods [1] is frequently studied by computing intercorrelations among ratings made by a single rater.

Gerberich,[2] in a study of the consistency of questionnaire results, sought to determine whether the length of time between administrations of a questionnaire would affect consistency in answers. When there was a one-day interval the results showed 91 per cent consistency; when there was a seven-day interval the results showed a 76 per cent consistency; and for the ten-day interval, 73 per cent. The decline in percentage of consistency with lapse of time is attributed to carry over through memorization. The trend appears to follow the curve of retention. Factual questions showed less consistency than those involving attitudes or introspective data.

On the basis of his findings together with an evaluation of questionnaire research data, Gerberich makes a number of observations regarding the consistency of questionnaire answers. The belief that factual information is more reliable than attitudinal or introspective data is not substantiated by his findings. On the contrary, his data show that the reverse is true insofar as we judge accuracy by consistency. But Gerberich hastens to suggest that accuracy and consistency are not necessarily synonymous. He suggests that further research be conducted to determine the accuracy of questionnaire responses as well as the consistency of responses such as interviews and autobiographical or other undirected forms of communication.

[1] See for example, L. W. Richard and W. Ellington, "Objectivity in the evaluation of personality," *Journal of Experimental Education*, 1942, *10:* 228–237.

[2] J. B. Gerberich, "A study of the consistency of informant responses to questions in a questionnaire," *Journal of Educational Psychology*, 1947, *38:* 299–307.

Discrimination. Although discrimination is intimately related to validity and reliability, its importance deserves special consideration. Usually when we employ a research instrument it is for the purpose of comparing and contrasting two or more things (objects, groups, or individuals). If the instrument does not show any differences between or among these things, we of course still have a finding, though for our purpose it is often a trivial one.

Suppose a college admissions officer is trying to differentiate between two groups of persons: those who will succeed in his institution and those who will fail. If the test that he uses yields approximately the same score for every candidate and if a proportion of these individals later make unsatisfactory records at his institution, then the instrument will have failed to discriminate. Similarly, if there are high and low scores (say, on a speed-of-tapping test which bears no relationship to subsequent scholastic achievement), then discrimination has not been obtained. In the former instance, invalidity may have resulted from lack of variability; in the latter, the test may have been reliable but it did not "tap" the proper function.

Even when measuring the height of a single boy we are concerned with several kinds of discrimination: contrasting his height now with his height a year ago, comparing him in this respect with other boys of the same age, attempting to estimate how tall he will be when full grown. In order to make sure that proper discriminations are obtained a reliable instrument is necessary. The third situation calls for, in addition, a substantial correlation between present and ultimate height.

Norms and sampling. The importance of norms and of certain sampling procedures has already been indirectly considered in our discussion of validity and reliability. These checks on validity and reliability involve concepts that are dependent upon certain considerations in sampling the data to be used in standardizing research instruments. Without appropriate sampling procedures, including representative subjects of the group for which the instrument is to be used, there could be no satisfactory means of estimating validity and reliability. The extent to which norms may be regarded as satisfactory depends upon (1) whether the group on which the instrument was standardized is representative of the group on which the instrument is to be used, and (2) whether the number of cases in the group on which the instrument was standardized is sufficiently representative.

When evaluating an instrument for use with a group we should be most concerned with the appropriateness of the norms for the

particular group upon which the instrument is to be used. If, for example, we find that the Gamin Personality Inventory (Guilford) has been standardized on college students (men and women) and we wish to use the inventory for measuring lieutenant colonels in the army whose mean age is approximately 38 years, norms should be established on the basis of the army group. Usually in adjustment and interest inventories, norms for several groups are provided; for example, adults, college students, and high school pupils. Analysis also is often made according to sex. An individual's score should be compared with the norm for the group of which he is a member.

A number of writers [1] have shown that a systematic bias may be introduced through sampling procedures. Various factors to be considered in selecting a population on which an instrument is to be standardized include age, sex, geographical location, education, nationality, and race—all of which may influence the effectiveness of a particular method of sampling.

In addition to making reference to norms that may accompany a standardized instrument, it is often desirable to derive local norms for particular situations. The derived norms for a particular test, for example, may be too high or too low for a certain population. A good example of the use of both national and local norms on the American Council on Education tests is found at the Herzl Junior College (Chicago). The norms for entrance to four-year colleges and universities are used in counseling students who may wish to transfer to higher institutions upon the completion of junior college. They are used in helping to solve the curriculum problem—that of determining the extent to which the preparatory and the terminal functions should be stressed. The local norms are also used in dealing with individual differences. School superintendents should be aware of the need of interpreting both local and national norms in relation to curriculum objectives, and variations in school populations from one school to another within the same system.

SUMMARY

Procedure in the evaluation of research instruments will operate in two ways: (1) development of refined criteria for appraising the large number of instruments now available; and (2) construction of instruments, when needed, in the light of such criteria. Since

[1] See for example, E. S. Marks, "Selective sampling in psychological research," *Psychological Bulletin*, 1947, *44*: 267–275.

there are now available a large number of research instruments their evaluation is a problem in itself. The personal interviews, questionnaires, and inventories have been extensively used as a means of gathering data; their reliability and validity as instruments of research should be established. Rating methods have been a basis for appraisal in almost every field of education; they need to be studied adequately for accuracy and adaptability to research problems. When instruments are not available some must be constructed skillfully and evaluated critically prior to their use in an investigation. After reliability has been established, there is the further problem of determining the validity of instruments, which is a more difficult task inasmuch as we are never sure that the criteria used for its establishment are in themselves adequate.

CHAPTER V

The Description and Appraisal of Status

The type of appraisal to be considered in this chapter has been variously named. It is sometimes referred to as a normative survey, as a descriptive investigation, or as a status study. The purpose of such studies is to develop an adequate description of the status of some phenomenon. The appraisal of status is usually accomplished by comparing the status of the phenomenon under investigation with expectancies, as expressed in objectives, standards, or criteria; or with norms obtained from studies of similar phenomena. The ultimate concern is not usually with status *per se*, but with the adequacy of status, once this has been ascertained. It will help the reader to keep in mind such words as *description*, *status*, *norms*, *standards*, and *criteria* as the discussion proceeds.

THE KINDS OF PROBLEMS TO BE DISCUSSED IN THIS CHAPTER

There are many occasions upon which one may need to describe and appraise status. They include all attempts to determine the adequacy of the educational product. They also include appraisals of the conditions influencing educational outcomes. Usually, one is not satisfied to know merely the adequacy of the product; one may also wish to know the status of the many conditions that limit or facilitate educational outcomes. There are many such conditions that may limit and facilitate pupil growth and achievement, the matter of our ultimate concern in this volume. Some of the conditions reside in the setting for learning: the physical plant, financial support, administrative organization, the social structure of the

school, home background, and community resources. Some reside in the educational personnel: teachers, administrators, and supervisors. And others reside in the pupils themselves: their general intelligence, aptitudes, interests, attitudes, knowledge, skills, and acquired behavior patterns. These conditions are well established in the thinking of most educators and may be subjected to systematic appraisal.

Frequently, too, one must describe and appraise various kinds of educational processes: the learning process as it relates to many kinds of learning; the teaching process as it relates to subjects, activities, and outcomes; and administrative processes as they relate to numerous aspects of the educational program. Appraisals of processes may arise from a desire to know the kinds and amounts of training and experience of the teacher personnel; the kinds and amounts of retardation or of acceleration among elementary school pupils; the stated causes and frequency of turnover in the administrative personnel; the amount of financial support accorded different educational services; the content of textbooks in various subject areas; the kind and amount of instructional materials; supplies and equipment provided for the different subject areas and levels of instruction; and the types and costs of school buildings in cities of varying size. There are many such studies reported in educational literature. To define further the subject matter of this chapter a list of studies is provided at the end of the book.

To know status is frequently important. Planning, putting plans into operation, and appraising results are important educational operations. Without knowledge of status and its adequacy there is much working in the dark. Not only is a knowledge of status important in and of itself, but as a foundation for the interpretation of many other kinds of data, such as the data for experimental and correlational studies, the data for historical and developmental studies, and the data for many kinds of comparative studies to be discussed in a later chapter.

Status studies may be qualitative or quantitative. At one level, descriptions of status may consist of naming and defining constituents, elements, or aspects of various phenomena, such as the qualities of a good teacher; characteristic educational practices, acceptable or non-acceptable; the activities of the school personnel or the pupils; and the units of subject matter or experience that may constitute a curriculum. At another level status studies may involve ascertaining the amounts of constituents, elements, or characteristics. At one level one's concern may be merely with the presence or absence of certain elements, attributes, or constituents. At

another level one's concern may be with the amounts of each element, attribute, or constituent.

Descriptions may be made through the use of either verbal or mathematical symbols. Many behavior descriptions currently found in the literature of education are of the verbal kind. That is, the symbolism is verbal; other descriptions are made in terms of countable units, or mathematical symbols. It is sometimes said that qualitative studies are verbal whereas quantitative studies are mathematical. This is not entirely correct, since there are many verbal symbols in our language that may be used to indicate quantity, such as the words: *few, many; frequently, infrequently; seldom, never, always; large, small; high, low; near, far; heavy, light;* and *slow, rapid.* As the thinking in any particular area of research becomes more refined the tendency is, however, to substitute mathematical symbols for verbal symbols; to do so, one must have countable units of measurement.

Data may be variate or nonvariate. Mathematical data may be of two types: variate or nonvariate. Although there is considerable overlapping in these two categories, nonvariate data are derived from tabulating and counting the occurrence or nonoccurrence of elements, constituents, or attributes; variate data are derived from measuring the amounts of various attributes. To count, there must be discernible wholes, categories, or units—such, for example, as the sex and geographic distribution of persons in a census report; or the presence or absence of certain characteristics of persons, situations, or actions, taken as a whole. Nonvariate data are frequently referred to as categorized data. In measurement, however, one counts also (that is, one counts units of measurements: inches, pounds, degrees), but the end result is an ordered statement of the amount of some element, constituent, or aspect of something, such as the height of sixteen-year-old boys, the mental age of twelfth-year high school pupils, the reading age of ninth-grade science pupils, and the speed of computation of third-grade pupils performing certain selected arithmetical exercises.

Mathematical analyses may be descriptive or inferential. In the field of statistical analysis there are two ways of treating data: to one of these ways we assign the name *descriptive statistics* and to the other the name *inferential* or *sampling statistics.* In descriptive statistics one's concern is with certain characteristics of some immediately available group of objects such as a class of pupils. In sampling statistics one's concern is not so much with the immediate group as in the use of information concerning the group to draw inferences regarding some larger population of which the group

is a part. To draw statistical inferences, however, one does not begin with just any immediately available group but with a carefully drawn sample. In descriptive statistics the goal is accurate information concerning the group at hand. When one's concern is with a larger population where it is not feasible to make a direct examination of all of its members, one will turn to *sampling* research described in Ch. VI. The sampling survey may involve types of problems discussed in this chapter, but most importantly it provides a means of dealing with larger groups according to the interests of the investigator. In inferential statistics the goal is accurate information concerning some more inclusive population, about whose members it would be impossible or impracticable to collect all the information that one desires. Descriptive statistics consist principally in the reduction of data about groups to more manageable wholes through use of tables, graphs, and numerical calculations; inferential statistics may do the same but its concern is with populations.

The purpose of this chapter is to discuss methods and techniques by which one may ascertain the status and adequacy of some educational situation. The data may be qualitative or quantitative; mathematical or non-mathematical, descriptive or inferential. The ultimate goal is valid and reliable evaluations of the products, processes, and conditions that underlie or result from action programs in the field of education. To make such judgments it is necessary to compare carefully the status of the phenomenon under investigation with that of similar objects of the same category, with norms, or with carefully validated standards or criteria.

VERBAL DESCRIPTIONS AND APPRAISALS OF STATUS

Investigations of the type under consideration in this section are for the most part observational studies that employ verbal symbols, as the primary means of recording, analyzing, and summarizing data. They may relate to any one of a number of aspects of the educational program, such as processes: time studies, activity analyses, and sequential analyses; educational products such as social adjustment, personality or skill in the communicative arts; or the conditions believed to limit or facilitate educational outcomes: buildings, supplies, equipment, administrative organization; and teacher-pupil relationships. Process studies will be described first.

Process studies. Although process studies are of many kinds the two most common are those describing operational sequences. One attempts to ascertain the order in which various part activities are

performed: for example, the steps in problem solving, the steps in initiating a learning experience, the sequence of events in group action, or the experiences leading to various kinds of maladjustments. In a discrete-phase study one attempts to ascertain without regard to sequence what is done. Studies of the methods and techniques of doing all sorts of things associated with the educational program may be classified as discrete phase studies. Two illustrations of process studies follow: one involves a very complex socio-educational activity, namely curriculum building and the other the reading process.

Changing the curriculum. An illustration of a nonquantitative descriptive study of process will be found in Miel's *Changing the Curriculum.*[1] The book is primarily a report of findings or generalizations; it is based, however, upon a study of a number of particular cases of curriculum making. Examples of these concrete materials are given in the appendices, and include such items as excerpts from professional logs; basic assumptions for curriculum planning in the public schools of Philadelphia; and curriculum development in Maine. The report of the Maine project discussed and executed such items as the following:

a) The development of a viewpoint on aims and purposes.
b) An initial conference to test opinion and set the stage for a democratic program.
c) Superintendents and teachers were asked for their views.
d) The normal schools offered their contributions.
e) Social study groups and regional conferences were organized.
f) Summer workshops were organized.
g) Bulletins appeared.
h) Social communities were consulted.
i) Co-operatively planned local programs arose, etc.

The author outlines three guarantees that she considers essential in evaluating a social process, namely: (*a*) the guarantee of security, (*b*) the guarantee of individual and group growth, and (*c*) the guarantee of accomplishment. In order that the process of deliberate social change may include the guarantee of security, growth and accomplishment, four groups of factors should be recognized and controlled, namely: (1) the motivation of the persons on whom change depends; (2) conditions of effective group endeavor; (3) the extent of social invention; and (4) the amount and quality of leadership present. Among the suggestions proposed by the author for

[1] Alice Miel, *Changing the Curriculum: A Social Process* (New York: D. Appleton-Century-Crofts, Inc., 1946), p. 242.

the guidance of educational leaders in curriculum making are the following:

1) Provide for changes.
2) Discover an adequate process.
3) Respect the principle of gradualism and rapidity.
4) Recognize the importance of values.
5) Capitalize upon complaints.
6) Recognize the need for self-set goals.
7) Set up a simple and functional internal organization.
8) Strive for a condition of diversity within unity.
9) Help participants operate increasingly on the basis of new knowledge.
10) Bring the available and needed expertness to bear upon the situation.
11) Make constructive use of communication.
12) Practice and extend techniques of group action.
13) Build constructive social power.
14) Regard authority as something residing in a working group.
15) Develop expertness in techniques of group action.
16) Generate as much leadership in others as possible.
17) Become increasingly familiar with principles of human development.

The author concludes as follows:

> Much remains to be learned in the field of educational leadership and many problems can be solved only in the light of a given set of circumstances. Among them are several raised in the foregoing discussion: (1) the most desirable ways of enlisting the initial interest of teachers, learners, and community adults in a process of curriculum change; (2) the amount of diversity which can be tolerated comfortably in a school or school system; (3) ways to help principals, supervisors, and other specialists in the school system to find creative roles; (4) the amount of freedom desirable for all concerned with curriculum development at different stages of the process. It is to the solving of all such problems relating to curriculum change that the status leader in education must address himself in the years ahead. The patient and painstaking forging of an adequate process of curriculum change for each American school is a task that will require sound and imaginative leadership of groups that are constantly improving their values and their techniques.

Another example of curriculum research of this type will be found in Emans' report [1] upon curriculum making in Dane County, Wisconsin.

[1] Lester M. Emans, "In-service education of teachers through co-operative curriculum study," *Journal of Educational Research*, 1948, 41:695–702.

Studies of the reading process. Many studies have been made of the reading process, a considerable number of which have employed elaborate instrumentation such as Buswell's [1] early studies of eye movements. Process studies of reading are of two kinds: (1) those relating to motor processes and (2) those relating to mental processes. The motor processes may be classified as (*a*) visual, (*b*) vocal, (*c*) extraneous; and (*d*) mental processes, such as word perception, apprehension of meaning, and related processes. In studies of perception it has been found, for example, that the eyes move along the lines of print in a series of short movements and pauses. The number of words perceived at each fixation varies widely among readers, particularly among good and poor readers. There are three distinct views of the nature of the process by which words are perceived: (1) one group of persons maintains that the context of what is read provides mental set and arouses the associations essential to word recognition; (2) another group maintains that the word is the unit of recognition and its form provides the characteristics by which it is recognized; (3) a third group attaches primary importance to the letters of which words are composed.[2]

Vernon [3] identifies four steps in the perception of words: (1) vaguely perceived form of contour with (2) certain dominating or specific parts, which (3) stimulate auditory or kinesthetic imagery, and (4) arouse meaning. The generalizations indicated above are illustrative of many that have grown out of research in this area. Each of these researches grows out of a carefully defined procedure: a carefully stated problem; a description of what is to be observed or studied; a justification of the data-gathering devices to be employed, including appropriate instrumentation; the major characteristics of the group studied; the controls employed; and other considerations that make possible a precise interpretation of the findings. Only by careful study of many examples of research such as those listed in the bibliographies cited in the references can the beginning research worker develop the insights, specific knowledges, and skills necessary for high level research and appraisal in this area. The studies here referred to are primarily observational and descriptive.

[1] Guy T. Buswell, *Fundamental Reading Habits: Study of Their Development*, Supplementary Educational Monographs, No. 21 (Chicago: University of Chicago Press, 1922), 150 pp.

[2] Walter S. Monroe and others, *Encyclopaedia of Educational Research*, Revised Edition (New York: The Macmillan Co., 1950), p. 976.

[3] M. D. Vernon, *The Experimental Study of Reading* (London: Cambridge University Press, 1931), 190 pp.

The appraisal of products. Frequently the description and appraisal of status relate to educational products. The commission on the relation of schools and colleges established by the Progressive Education Association for the Eight-Year Study of Secondary Education was interested, for example, in the appraisal of the complex forms of pupil behavior that result from the modern school. Many hours of careful planning preceded the data-collecting phases of this study. Educational purposes, assumptions, and evidences of pupil growth were carefully defined by various groups of experts. Many quantitative measures were developed and used for collecting information about the aspects of school education with which the commission was concerned. They also developed many nonquantitative descriptive techniques. It is these nonquantitative descriptive techniques with which we are concerned here. One of these related to techniques employed in collecting and recording information about pupil behavior. The committee states its purpose with reference to this project as follows: [1]

> It will be clear from the material itself that the method of studying pupils devised by the committee depends on the studying of descriptions of the different kinds of behavior that are likely to be observed in relation to the characteristics chosen. The descriptions made by the committee are designed to define what might be called type or classifications of behavior in terms of each characteristic. The use of carefully worded standard definitions in place of teachers' own wordings is intended to bring about a more nearly common understanding of the characteristics themselves and of persons described.
>
> In general, all teachers having opportunity to know a pupil would be expected to describe him by the use of this material. The combined reports, which would appear on the Behavior Description Card, would show the pupil's most common behavior, as well as the range of behavior under different conditions.

The Behavior Description Card referred to above consists of (1) a listing of characteristics and the description of the classifications set up under each, (2) space for data, (3) a key system for use in recording the judgment of teachers, and (4) space for general comment.

The qualities about which data are to be collected are: responsibility-dependability, creativeness and imagination, influence, inquiring mind, open-mindedness, power and habit of analysis, social

[1] Eugene R. Smith, Ralph W. Tyler, and the Evaluation Staff, *Appraising and Recording Student Progress* (New York: Harper and Brothers, 1942), pp. 473–4.

TABLE 5. Excerpt from the Form on Admissions (Commission of the Relation of School and College)

NOTES: The following characterizations are descriptions. They are not ratings. Supplementary or alternative descriptions will be found under "General Comment." M (Mode) followed by a number indicates the most common behavior of the pupil as judged by that number of teachers. Significant deviation from the common behavior is shown by the name of a subject-field or other pupil-teacher relationships in the appropriate space.

Work Habits	Highly Effective	Adequate	Promising	Limited	Ineffective
		3–M–3		Music	
Serious	Purposeful	Limited	Potential	Unreliable	Vague
Purpose	M–4	3	Art–Music		Phys. Ed.
Responsibility Dependability	Responsible and resourceful	Conscientious	Generally Dependable	Selectively Dependable	Irresponsible Unreliable
	Math.–Home Room	M–4	Art–Music		Phys. Ed.
Creativeness and	General	Specific	Promising	Limited	Imitative–Unimaginative
Imagination		Math.–Nat. Sc.		M–4	Music–Phys. Ed.
Influence	Controlling	Contributing	Varying	Co-operating	Passive
		M–4	Nat. Sc.–Eng.	Music	Phys. Ed.
Inquiring	General	Specific	Limited	Directed	Unresponsive
Mind	Math.	Nat. Sc.–Eng.	M–4	Phys. Ed.	
Power and Habit	Highly Analytical	Incomplete	Irregular	Undeveloped	Limited Passive Unreasoning
of Analysis		Math.–Nat. Sc.	M–4		Music–Phys. Ed.
Concern for Others	Generally Concerned	Selected concerned	Personal	Inactive	Unconcerned
		M–6	Music		Phys. Ed.
Personal	Secure	Uncertain	Neutral	Withdrawn	Not Accepted
Adjustment	4–M–4				Phys. Ed.
Self-Reliance	High		Usual	Low	
Aesthetic Appreciation	High		Usual	Low	

NOTES: Number of faculty members reporting? Nine on what are the descriptions based?
Definitions of these headings made by records and reports committed? Yes
On anecdotal records? ____ Yes No
What other basis? ____

concern, emotional responsiveness, serious purpose, self-reliance, aesthetic appreciation, social adjustability, and work habits.

An excerpt from these data taken from the form sent to the committee on admissions is reproduced in Table 5.

The analysis is followed by general comment as follows:

GENERAL COMMENT (Made by) Principal

(The following information amplifies the description of the candidate. It should include the characteristics under "Behavior Description" if the table above is not used, and should add anything important about family background, possible financial needs, and accomplishment in terms of special objectives of the school.)

Mr. and Mrs. Doe are good members of the community and are cooperative in their relations with the school. They expect to be able to meet the cost of John's college education.

John has been somewhat handicapped by a severe illness at the beginning of his secondary school course, but he is now increasingly showing power, particularly in mathematics and related work. He wishes two years in liberal arts, probably followed by preparation for engineering, and we believe this plan to be a good one.

While John is shy in some situations, and he avoids physical activity altogether too much, he is on the whole a good member of a group and a boy with promise for college success and for later usefulness.

There are many such measurement and appraisal techniques to be found in the literature of education. The illustration presented is of a complex educational product. Any number of illustrations might have been chosen from the field of survey testing.

The appraisal of conditions thought to limit or facilitate educational outcomes.[1] Up to this point our discussion has concerned educational outcomes. Frequently one needs for his purposes not merely an evaluation of products of an educational program but appraisal of the conditions that underlie or accompany various outcomes; such, for example, as the social and physical setting for learning, the personnel, or the available resources. There are numerous such conditions. They may all exert a profound and continuing influence upon educational programs.

Such studies proceed from well-defined purposes and description of the group to be studied. The data may be variously collected through the use of tests, interviews, questionnaires, inventories, observations and various mechanical devices, and rec-

[1] A. S. Barr, William H. Burton, and Leo J. Brueckner, *Supervision* (New York: D. Appleton-Century-Crofts, Inc., 1947).

ords. After the data are collected they are evaluated against norms of behavior or criteria. The criteria may be carefully formulated and validated, or they may be of the unwritten variety existing in one's mind. To obtain valid and reliable evaluations there must be good data and good criteria and standards in terms of which evaluations are made.

To illustrate such appraisals two examples have been chosen for discussion. One example involves a study of the social structure of classroom situations, and the other a study of the personal factors in teaching efficiency. The purpose of the investigator in each instance is to develop a verbal summary of certain conditions believed to limit or facilitate educational outcomes.

A study of the social structure of a classroom situation. Learning always takes place in a socio-physical setting. In teacher-centered education this setting as observed from the child's point of view is frequently overlooked or greatly underestimated. Many social pressures growing out of home-school-community settings influence the child. Within recent years a number of attempts have been made to derive more adequate information about these forces.[1]

In a study of group relationship one must select the aspects of the situation that merit investigation and devise suitable data-gathering instruments. The test situation, if it involves children, must provide opportunities that are meaningful and natural such as choosing companions under various conditions: for sitting together, working together on committees, or developing projects. The data may be collected through observations of behavior and interviews. In any case the choices must be real and not artificial or hypothetical. In analyzing the data one attempts to answer questions such as the following: How many reciprocated choices are made? Which members of the class are most in demand? Which members are ignored? Do boys and girls choose each other? Do pupils of the same age, race, socio-economic status, interests, and intellectual capacity choose each other? Although there are many ways of collecting such data, they are usually summarized in a sociogram such as that presented in Figure 2.

This particular sociogram is read by noting the lines that lead from one pupil to another.

The circle marked "Mary Jokin" in the lower left corner has three arrows (unreciprocated choices) running from it to Janet Toll (first),

[1] Helen Hall Jennings, *Sociometry in Group Relations: A Work Guide for Teachers* (Washington, D. C.: American Council on Education, 1948).

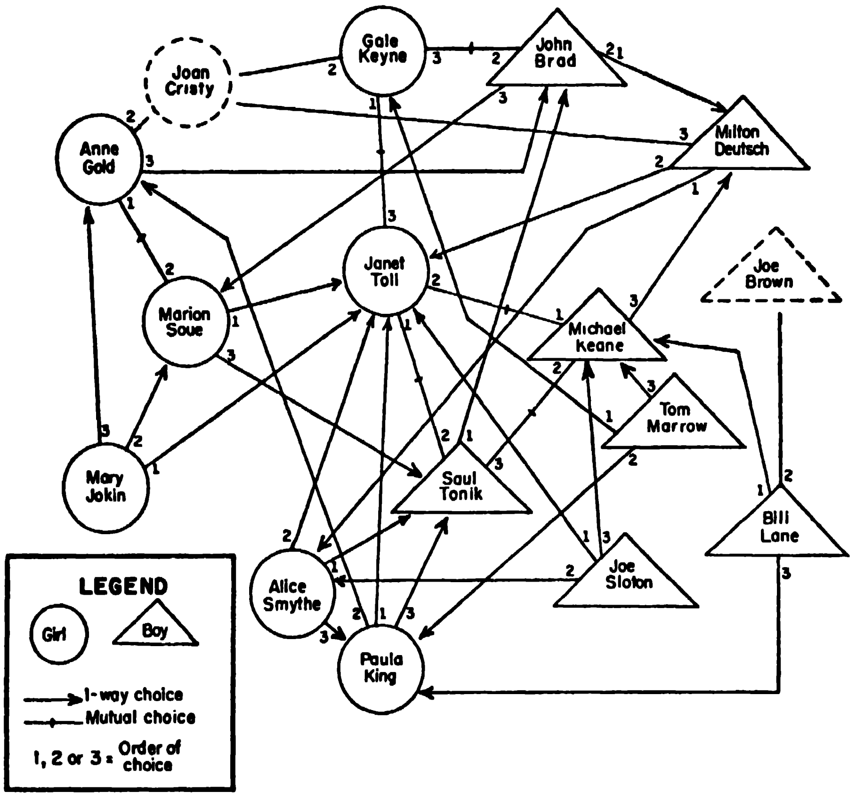
No. of Boys ________: No. of Girls ________ Class/Grade ____________

School ____________

City ____________

Date Given ____________

Test Question: ____



NOTE: For an absent boy or girl use the respective symbol dashed, leaving any choice line open-ended (see the case of Joe Brown in the above sociogram).

If rejections are obtained, the choice line may be made in dashes or in a different color.

Whenever a direct line from chooser to chosen cannot be drawn without crossing through the symbol for another individual, the line should be drawn with an elbow, as in the case of Bill Lane to Paula King.

FIGURE 2. Sample sociogram form.

Marion Soue (second), and Anne Gold (third). There are no arrows pointing at Mary; she has not received a single choice. Looking at Janet, we find that her first choice is a boy, Saul Tonik, and it is reciprocated as Saul's second possibility; her second choice is another boy, Michael Keane, and is also reciprocated—she is first on Michael's list; her third choice is again reciprocated by a girl this time, Gale Keyne, and represents another first choice. In addition, there are six arrows pointing at Janet, coming from two boys and four girls; moreover, four of these are first choices and the others second ones. Everyday life in the classroom must obviously be very different for Janet and Mary!

On looking further at the sociogram for any other patterns, a mutual choice will be discovered between Saul and Michael—the two boys who chose Janet. Here is a triangle established on the basis of mutual choices on all three sides. Another pair relation exists between Janet's friend Gale, and John Brad, and still another between Anne and Marion, both of whom had been vainly chosen by Mary. Other sociograms might show additional triangles or squares or pentagons of mutual choices, dividing the class into two or more clearly defined groups. If these patterns of relation are completely self-contained with no arrows or lines running between them, it means that friendship goes by cliques. But that would be an extreme situation. The most frequently encountered pattern within the over-all network is a sort of string or chain of one-way choices; there is a twisting one on the sample sociogram, running from Mary to Anne to John to Milton to Alice to Janet. In the primary grades such chains occur very often and are sometimes quite long. Usually there are few mutual choices in the sociogram until the third grade.[1]

In order to understand fully the social structure of a classroom situation it is necessary to develop many such sociograms. To interpret a sociogram it is necessary to know many facts about children. Various forces may be thought of as influencing a child in a field. The behavior of the child is determined by his internal conditions and the nature of the field forces. Effective appraisals of status cannot be made without consideration of these many forces.

A study of personal factors in teaching efficiency. The data referred to here are part of those collected in connection with the Beloit, Wisconsin, study of teaching efficiency. Many aspects of teaching efficiency were studied. We are concerned here only with the personal factors believed to condition efficiency.

In making such a study one must first formulate ideas about what to look for. These are found in one's own personal experiences, in the personal experiences of others, and in the published reports of systematic research. All of these sources of ideas were

[1] *Op. cit.*

used in this study, but in the final stages of the investigation a very careful examination was made of four investigations: (1) the Barr and Emans [1] analysis of 209 teacher rating scales; (2) the Charters and Waples [2] *Commonwealth Teacher Training Study,* particularly those sections relating to personal fitness; (3) Barr,[3] a summary of researches relating to the measurement and prediction of teaching efficiency, and (4) Cattell's [4] *Description and Measurement of Personality.*

From these sources a tentative list of traits of personality was secured. These were then submitted to a seminar group for examination and systematic checking. This check involved the rereading of many statements about the personality of teachers and the holding of many group conferences on terminology. The list of trait names (and definitions) developed in this seminar was next submitted to a representative group of Beloit teachers and administrators and finally to the entire teaching and administrative staff for study and revision. Some trait names were eliminated, some added, and some combined with others. The list finally accepted for further study is given below:

1) ATTRACTIVENESS—Dress, physique, absence of defects, personal magnetism, neatness, cleanliness, posture, personal charm, appearance.

2) CONSIDERATENESS—Concern for the feelings and wellbeing of others. Sympathy, understanding, unselfishness, patience, helpfulness.

3) DRIVE—Habitual readiness for effective action. Force, vigor, energy, eagerness to succeed, ambition, motivation, vitality, endurance.

4) RESOURCEFULNESS—Capacity for approaching things in a novel manner; initiative; originality.

5) REFINEMENT—Good taste, modesty, morality, conventionality, culture, polish, well-readness.

6) CO-OPERATIVENESS—Friendliness, easy goingness, geniality, generousness, adaptability, flexibility, responsiveness, and warm-heartedness, unselfishness, charitableness.

7) RELIABILITY—Accuracy, dependability, honesty, punctuality, responsibility, conscientiousness, painstakingness, trustworthiness, sincerity.

8) EMOTIONAL STABILITY—Realism in facing life's problems, freedom from emotional upsets; constancy; poise; self-control.

9) BUOYANCY—Optimism, enthusiasm, cheerfulness, gregariousness,

[1] A. S. Barr and Lester M. Emans, "What qualities are prerequisite to success in teaching," *The Nation's Schools,* VI (September, 1930) p. 26.

[2] W. W. Charters and Douglas Waples, *The Commonwealth Teacher Training Study* (Chicago: The University of Chicago Press, 1929), pp. 14–19; 50–56.

[3] A. S. Barr, "A summary of researches relating to the measurement and prediction of teaching efficiency," *Journal of Experimental Education* (September, 1948).

[4] Raymond B. Cattell, *The Description and Measurement of Personality* (Yonkers, New York: World Book Company, 1946).

fluency, talkativeness, sense of humor, pleasantness, carefreeness, vivaciousness, alertness, animation, idealism, articulation, wittiness.

10) OBJECTIVITY—Fairness, impartiality, open-mindedness, freedom from prejudice, sense of evidence.

11) ADAPTABILITY—Ability to adjust to unforeseen circumstances, individual differences in persons, and unusual social climates.

12) INTEGRATION—Organization; activity-willed purposeful behavior; expeditiousness of action.

Through the co-operative activities of all the teachers and administrative officers each quality was ultimately defined in terms of observable behavior. The analysis for two of these qualities will suffice to illustrate the decisions reached.

OBJECTIVITY

( ) 1. Does the teacher have a good sense of evidence and use it at all times (that is, does he keep an open mind in the absence of complete and accurate information)?

( ) 2. Does the teacher allow the facts of the situation to determine action rather than personal feelings (that is, makes a factual rather than an emotional approach to problems and situations)?

( ) 3. Has the teacher the ability to see all sides of a problem or situation (that is, gets all around a subject; sees the whole rather than limited aspects)?

( ) 4. Is the teacher impersonal in his comments to, criticism of, and suggestions for pupils, co-workers, and members of the community?

( ) 5. Is the teacher free from favoritism, prejudice, and preconceived ideas in dealing with pupils, co-workers, and members of the community?

BUOYANCY

( ) 1. Is the teacher able to meet disappointment with hope rather than despair?

( ) 2. Is the teacher able to find good in most persons and situations?

( ) 3. Is the teacher free from moodiness and depressiveness?

( ) 4. Does the teacher show a lively interest in his/her work?

( ) 5. Does the teacher have a happy relaxed disposition free from tensions?

( ) 6. Is there an absence of excessive complaints and griping?

( ) 7. Has the teacher a good sense of humor?

( ) 8. Is the teacher expressive, in speech and actions?

( ) 9. Has the teacher a cheerful disposition (not grouchy, cynical, or disillusioned)?

The separate items, were scored as satisfactory, unsatisfactory or uncertain, with suitable annotations. The same procedure was em-

ployed in recording judgments about the teacher's general fitness or efficiency.

There are many problems common to the types of studies here discussed. First, there is the problem of semantics, present in all research, but characteristically so in this type of study. Secondly, there is the problem of clearly distinguishing between data and inferences drawn from data; and finally the problem of categorizing.

The semantics problem in qualitative studies. Verbal symbols must always be used with great care. They possess different meanings for different persons at different times and in different contexts. When one speaks of qualities, as we do in most appraisal studies, they may mean almost anything, depending upon the accuracy and objectivity with which they are defined. This difficulty is evident in both observation and interrogation, as in interviews and questionnaires. Questions may be variously interpreted by different respondents and the data so collected meaningless. Not only is there difficulty in defining terms in such a way that we can be certain that they are present or absent, but there is also difficulty in answering questions of extent and degree. Because of temperamental differences in individuals and their habits, words such as *frequently*, *strongly*, *vigorously*, *inadequately*, and *slightly*, may have different meanings. "Frequently," for example may mean almost anything. The difficulties cited are only illustrative of the many that one may experience in defining terms, establishing categories, and summarizing results. By the time one reaches the concluding aspects of studies such as those cited, the degree of error may have grown to tremendous proportions.

Distinguishing between fact and inference. Only in a general way is it useful to draw a distinction between fact and inference. Some distinction, however, seems necessary. Almost any observable object, as for example, a nail in a board, may be observed at a close range. Let us make a statement of fact about it. We may say, for example, that it is a fact that the board has a nail in it. But is this a fact or an inference drawn from sense data present in the observer's mind? If, however, two or more individuals observe the same phenomenon simultaneously or under comparable conditions and agree upon the presence or absence of this phenomenon, we may and frequently do call the object, or action, a fact. The matter becomes more complex, however, when we enter the realm of value judgments. Two or more observers of an athletic contest may agree upon the occurrence or nonoccurrence of certain acts, such as slugging, tripping, or pushing, but not upon legitimate inferences

about them. They may not agree, for example, upon which team was the more sportsmanlike, skilled, or spirited. Much that is said to have been observed is merely a reflection of the observer's preconceived ideas as to what should take place and the values that he attaches to such events.

Requirements in developing verbal descriptions and appraisals of status. Although not all of the important aspects of attempts at verbal descriptions and appraisals of status can be cited in this brief resumé, certain ones, to be further elaborated later, may be recalled, as follows:

1) There should be a clear statement of purposes, objectives, or goals.

2) There should be a careful statement of what is to be described or appraised, whether it be products, processes, or conditions for effective operation.

3) There should be a careful description of subjects studied whether these be persons, ideas, or inanimate physical objects.

4) There should be a description of how the data were collected, including information on the validity and reliability of the data-gathering devices employed.

5) There should be a careful discussion of how the data were categorized, summarized, and analyzed.

6) There should be some indication of the assumptions made and the criteria employed when value, judgment, or appraisal is made.

7) Terms must be carefully defined.

8) Inferences must grow out of the data collected and not preconceived ideas.

MATHEMATICAL DESCRIPTIONS AND APPRAISALS OF STATUS

Nonvariate mathematical studies of status. Early in this chapter we drew a distinction between qualitative and quantitative studies. We classified the large numbers of qualitative studies found in the literature into two categories; namely, (*a*) those employing verbal symbols, in recording, analyzing, and summarizing data, and (*b*) those using mathematical symbols for such purposes. Studies employing mathematical symbols may be further classified into those employing nonvariate data and those employing variate data. We shall now discuss some of the problems associated with nonvariate mathematical studies of status.

Many studies lend themselves to this type of investigation. Almost all aspects of the educational program may be subjected to

TABLE 6. Comparison of Mean Frequencies per Child of the Contacts of Teacher 2-D with Two Different Groups of Second-Grade Children in Consecutive Years (*observation time:* two hours per child).

Type of Teacher Contact		Teacher 2-D38 N = 29 Mean	S.D.	Teacher 2-D39 N = 28 Mean	S.D.	C.R.
		GROUP CONTACTS				
DC	Domination in conflict	7.9	4.4	12.0	5.5	3.2 *
DN–(1–8)	Domination with no conflict-directive	115.1	26.9	86.3	23.1	4.3
DN–(9–10)	Domination with no conflict-lecture method	93.7	31.9	102.0	27.6	1.1 *
DN	Domination with no conflict	208.8	49.7	188.3	39.9	1.7
DT	Domination in working together	3.0	2.0	1.6	1.5	3.0
	TOTAL DOMINATION	219.7	21.8	201.9	40.0	1.5
IN	Integration with no evidence of working together	38.8	11.7	23.7	10.4	5.2
IT	Integration in working together	2.2	2.0	0.7	0.9	3.7
	TOTAL INTEGRATION	41.1	11.7	24.4	10.5	5.7
Total Contacts		260.7	53.7	226.3	41.9	2.7
		INDIVIDUAL CONTACTS				
DC	Domination in conflict	7.9	6.9	4.6	3.3	2.3
DN–(1–8)	Domination with no conflict-directive	7.4	3.6	6.6	4.8	0.7
DN–(9–10)	Domination with no conflict lecture method	6.2	5.1	9.8	9.6	1.7 *
DN	Domination with no conflict	13.6	6.6	16.4	13.7	1.0 *
DT	Domination in working together	3.7	2.4	7.8	5.2	3.8 *
	TOTAL DOMINATION	25.2	12.6	28.8	19.1	0.8 *
IN	Integration with no evidence of working together	2.3	2.7	1.6	1.4	1.3
IT	Integration in working together	3.3	2.6	4.7	4.5	1.4 *
	TOTAL INTEGRATION	5.7	4.0	6.3	4.8	0.5 *
Total Contacts		30.8	14.8	35.1	22.5	0.8 *

* Indicates that the mean per child for 1939 was larger.

qualitative analysis and the results expressed as frequencies. Teachers, pupils, buildings, items of equipment, and almost any other aspect of education may be counted. The studies of teacher and pupil relations by Anderson[1] and others illustrate this type of study. The authors were particularly concerned with pupil behavior under different types of teacher leadership. A comparison of the mean frequencies per child of the contact of teacher 2-D with two different groups of second-grade children in consecutive years (1938 and 1939) is presented in Table 6. The types of teacher contact are carefully defined and the frequencies with which each type of contact occurred in a two-hour observation of each of a number of children are recorded. The critical ratios for the differences secured in the contacts of the same teacher in consecutive years are presented in Table 6. The mean frequencies per child of the behavior of the two classes of pupils are presented in Table 7. The authors calculated numerous coefficients of correlations. Our immediate concern is with the qualitative part of the analysis. From the data presented the authors conclude:

A. With reference to the teacher

1) That teacher 2-D showed a significant increase in group domination in conflict situations; she also showed an increase in group domination lecture method contacts; all other types of group contacts showed a decrease.

2) That teacher 2-D showed a decrease in individual domination in conflict; her domination-lecture method showed an increase as did domination in working together; there were no statistically significant changes in the frequencies with which other types of contacts occurred.

B. With reference to the pupil

1) There were no significant differences between the two classes for the categories of *leaves seat, plays with foreign objects, attacks status of other child, answers spontaneously in recitation, fails to answer in recitation,* or any of the J-v categories of *voluntary social contribution.*

2) The children in 1939 showed significantly higher frequencies for the categories *nervous habits, looking up at seat work, undermined child-child contacts, total responses in recitation,* and in the total J-r categories of *social contributions in response* to questions or invitations from others. The children of 1939 showed increases representing differences that approached significance in *commands other child, dominates other children, seeks help and tells experience in response to others.*

[1] Harold H. Anderson and others, *Studies of Teachers Classroom Personalities, III: Follow-up Studies of the Effects of Behavior.* Applied Psychology Monographs, No. 11. Published for the American Psychological Assoc. (Stanford Univ. Press, Stanford University, California, 1946).

TABLE 7. Comparison of Mean Frequencies per Child of the Behavior of Two Different Groups of Second-Grade Children with Teacher 2-D in Consecutive Years (observation time: two hours per child).

Behavior		*Room 2-D38** *N = 29*		*Room 2-D39* *N = 28*		
		Mean	*S.D.*	*Mean*	*S.D.*	*C.R.*
N. H.	Nervous habits	44.1	17.6	89.1	30.3	6.8 *
L. Up	Looking up	34.8	10.1	54.1	20.5	4.5 *
L. Seat	Leaves seat	6.0	3.8	5.1	2.6	1.0
Undet.	Undetermined child-child contacts	63.4	31.3	94.9	33.6	3.7 *
F. O.	Foreign objects	1.7	2.1	2.2	2.7	0.7 *
P	Conforming	44.7	16.6	34.4	14.7	2.5
N–1	Commands other child	1.6	2.0	3.0	4.1	1.5 *
N–2	Attacks status of other child	0.0	0.2	0.4	0.7	0.6 *
N–(1–2)	Dominates other children	1.6	2.0	3.4	4.1	2.1 *
M	Nonconforming	8.4	8.1	2.1	2.0	4.2
L–1	Answers spontaneously	1.1	1.5	1.5	1.8	0.5 *
L–2	Holds up hand	11.7	91.8	23.5	12.7	5.4 *
L–3	Answers when called	5.2	4.0	13.2	9.8	5.8 *
L–4	Fails to answer	0.4	0.8	0.6	1.3	0.4 *
L–(1–4)	Response in recitation	18.3	12.7	38.9	21.8	4.3 *
K–1	Seeks help	3.1	2.8	4.3	4.4	1.3 *
K–3	Contributes to own problem	1.7	1.5	0.6	1.1	1.6
K–4	Contributes to other's problems	7.1	6.3	4.3	4.0	1.9
K–(3–4)	Contributes to own and other's problems	8.8	6.3	4.8	4.2	2.8
K–(1–4)	Problem solving	11.9	7.8	9.1	6.9	1.4
J–v Voluntary:						
J–v–1	Tells experience	1.2	3.9	2.3	4.1	0.9 *
J–v–3	Suggestions	0.3	0.7	0.8	3.3	0.6 *
J–v–5	Holds up hand	1.0	1.2	0.4	7.2	0.9
J–v–6	Appreciation	0.1	0.3	0.1	0.3	0.0 *
J–v(1–6)	Total voluntary social contributions	3.1	4.1	3.7	3.2	0.6 *
J–r Response:						
J–r–1	Tells experience	0.3	0.7	2.4	2.2	2.6 *
J–r–3	Suggestions	0.2	0.5	0.4	0.9	0.4 *
J–r–5	Holds up hand	0.6	0.7	1.3	3.0	0.9 *
J–r(1–6)	Total social contributions in response	1.4	1.3	4.4	3.5	4.3 *
J–(v & r)	Total social contributions	4.4	4.5	8.1	4.9	2.9 *

* Indicates that the mean for the children in 1939 was larger.

3) The children of 1939 showed significantly lower frequencies for *nonconforming;* and a decrease approaching significance for conforming and *problem solving.*

A survey of public opinion. Another example of the nonvariable qualitative study of status will be found in the public opinion poll. Massanari,[1] for example investigated the attitude of the people of

TABLE 8. Response, by Sample Group, of the Clinton and Cerro Gordo Samples to the Question: "In Your Opinion Does Combining School Districts Make Possible a Fairer Distribution of the Tax Burden?"

Sample Group in Prediction Study	*Per cent Responding "Yes" to Question*		*Per cent Responding "No" to Question*		*Per cent Responding "Uncertain" to Question*	
	Clinton	Cerro Gordo	Clinton	Cerro Gordo	Clinton	Cerro Gordo
Expressed intent to vote "For" new school district	70	85	4	2	26	13
Expressed intent to vote "Against" new school district	16	24	60	62	24	14
Said "Probably will vote but uncertain how"	12	39	12	3	76	58
Expressed no intent of voting	10	11	0	11	90	78

Illinois toward school district reorganization in selected areas. The data were collected by means of an interview questionnaire of 22 questions, such as the following:

7) In what size grade school, grades one to eight, do you think a modern educational program can be offered most economically?

a) Over 500 students
b) From 300 to 500 students
c) From 300 to 600 students, or
d) Less than 100 students
e) Don't know

10) Do you think that combining school districts will make possible more economies in expenditures? Tell me which of these answers best expresses your opinion:

[1] Karl L. Massanari, "Public opinion as related to the problem of school district reorganization in selected areas in Illinois,' *Journal of Experimental Education,* 1949, 17:389–458.

a) I am positive that it will
b) I think it might
c) I am uncertain
d) I doubt that it will
e) I am positive it will not

Some of the results are summarized in Table 8. The author reaches the following conclusions:

> It is possible on the basis of the evidence given in Section II, to make a dependable prediction of the outcome of a school district election provided that (*a*) opinions of all eligible registered voters are sampled; (*b*) the response is obtained during the ten-day period preceding the actual election, and (*c*) no major factors were introduced which would cause a sudden shift of opinion.
>
> The predictive technique was more accurate in estimating the per cent of favorable votes to be cast in the election than it was in estimating the per cent of eligible voters who would come to the polls.
>
> Evidence from the Clinton study suggested that those sample members who were most likely to go to the polls and vote were the people who returned their postal-card questionnaires.
>
> An analysis of the opinions held by Clinton and Cerro Cordo sample respondents of three related reorganization issues indicated that major differences of opinion existed between respondents who favored and those who opposed the establishment of a new school district.
>
> Voters generally thought that modern programs of education could be offered most economically in high schools with enrollments over 300 and in grade schools with enrollments over 100. A majority in Champaign and Urbana expressed opinion favorable to the merger of the two city school districts.
>
> When the response of the Champaign County respondents was broken down in terms of various background factors, the analysis indicated that the following subgroups were more liberal in their views about reorganization than were their counterparts: urban residents, younger age group, informed group, more educated group, males, upper economic group and school patrons.

The elements studied in this investigation are the opinions expressed upon the several questions asked. The results were summarized as frequencies and percentages.

Some aspects of nonvariate mathematical studies of status that need careful attention. After data have been collected, one's attention shifts to problems of tabulation and summarization. Two problems are of particular concern in this area: namely, (1) that of establishing appropriate categories for the classification of data; and (2) that of providing suitable summaries of data.

Establishing categories for the classification of data. In the absence of accurately established categories, it is possible to report a very large amount of inaccurate and misleading information. If this is not the case, we must be certain that our categories serve a useful purpose and are objectively defined. There are many aspects of most objects about which one can build categories, such as height, weight, color, or distance from some given point or geographic location. The grouping of books on one's study table can be made in many different ways and for different purposes. At one time one's purpose may be to gain some particular aesthetic effect; at another time to make certain books more readily accessible for use than others. The researcher must, through careful analysis of the situation develop useful categories. During his attempts to build such categories he might well ask himself, "Does this particular classification of attributes serve the purpose at hand?"

If the final tabulation of data is not to be misleading, the categories must be unambiguous: (*a*) they must be exhaustive, that is, include all possible divisions from some particular point of view and not a limited few divisions where there are many such categories; (*b*) they must be mutually exclusive, that is, not overlapping in a fashion such as to make the assignment of attributes to the several categories uncertain, and (*c*) they must be developed with reference to a single attribute, and not to a number of attributes simultaneously. The frequencies found in each category will depend upon the extent to which categories attempt to classify the data and their meaning.

One of the commonest errors in educational analysis is that of mixing categories. The products of learning are variously described in the literature of education as traits, behaviors, competencies, and mental controls. Each term is more or less ambiguous but where indiscriminately mixed in a single and possibly incomplete categorization the confusion is needlessly great. Not infrequently one finds tabulations employing categories such as cooperation, knowledge of subject matter, and the various competencies hopelessly mixed in a single categorization. The research worker will find it worth while to ask himself at the completion of each set of categories: "Have I maintained a single consistent point of view throughout in the establishment of categories?"

Questions may be of fact, attitude, or judgment. In some instances, as in tests and examinations, the interrogator presumably knows the answers; in others he desires to secure information that he does not possess. It is advantageous to remember that the "do know" and "do not know" answers to interviews, questionnaires, or inventory questions constitute a separate categorization in and

of themselves; the "do knows" can be subjected to further subcategorization as the situation may demand. Categories such as "like," "indifferent," and "dislike"; "like," "doubtful," and "dislike"; and "like," "do not know," and "dislike" need careful consideration.

Sometimes in the development of categories it is desirable to superimpose a variate upon a nonvariate classification. Such categorization necessitates two types of judgments: (*a*) the assignment of the item under consideration to its proper category of attributes, and then (*b*) some further division on the basis of frequency, intensity, or amount. One might, for example, attempt on the basis of an observation of behavior to answer the question "Is the individual considerate?" and then attempt further categorization on the basis of frequency: "often," "seldom," etc.

Much of the categorization in education is inadequate, showing the need for a wider acquaintance on the part of the research worker with the work of others, past and present. Studies based upon incomplete categorization are never satisfying and almost always require further research on the part of those who have a better grasp of the subject. In addition to being well read, possibly the best advice that one might give in this respect is to talk frequently with others about all proposed schemes of categorization.

Categories illustrated. Botts[1] devotes considerable space to the discussion of categories. In general she favors the a priori approach, assuming, if certain forms of behavior are being identified and recorded by various workers in the field that (1) these are commonly occurring forms of behavior, (2) they are readily identifiable, (3) they are significant.

Following this discussion, she reproduces an elaborate schedule of categories (five closely printed pages), employed in the St. George School, Toronto. The main divisions are:

Section I. General Categories
Section II. Motor Categories
Section III. Verbal Categories
Section IV. Adult-Child Categories

To illustrate further, there are eleven motor categories, with appropriate subdivisions and definitions, as follows: (1) random activity, (2) directed activity, (3) commands, (4) requests, (5) responses, (6) criticism, (7) repetition, (8) solitary play, (9) watching, (10) tires, and (11) smiles.

Botts[1] concludes with the following statement:

[1] Helen McM. Botts, *Method in Social Studies of Young Children*, University of Toronto Studies, Child Development Series, No. 1 (Toronto: The University of Toronto Press, 1933).

We have registered our conviction that in the choice of categories, observation is a better guide than logic, or to put it more adequately, observation should precede logic. The observer should bring a fresh, unbiased outlook to the work of observations, making the selection of significant and unambiguous forms of behavior a first task. The rearrangement of these forms into a systematic unity should be the last and not the first stage of category building. An ultimate objective in this connection would be a plan of categories applicable longitudinally over a sufficient age range to reveal significant stages in social development.

Another treatment of categories will be found in Murphy's[1] studies of social behavior. Her studies are based upon observations of behavior recorded in anecdotal form as episodes. The following episodes illustrate those for the conventional overtures of bright two-year-olds:

April 27
Davis showed a bunny wagon to Joyce. Davis said: "See, see? There's a bunny." He put it away, ran to Joyce, pointed to wagon, moved to another shelf. Ran to Joyce; laughed. Both went to piano. Davis pounded keys. Joyce imitated. Teacher took Davis and Joyce to bathroom.

April 27
Joyce approached Davis and said, "What are you doing?" Davis said, "Building." Tower fell. Davis laughed. Joyce said: "Go away," and left.

Mrs. Murphy was interested in the sympathetic behavior of children. The stimuli eliciting sympathetic responses among preschool children were classified as those arising from: *physical cause*—(1) accident, (2) attacked by child, (3) physical discomforts. *Mental distress*—(1) toy snatched or threatened, (2) play hampered or intruded upon, (3) disciplined, (4) separated from mother, father, etc., and (5) fear. *Emotional expression without knowledge of stimulus*—(1) crying, (2) holds hands over face, and (3) pained expression-inhibition of crying. *Evidence of injury without evidence of present pain, sore lip, mercurochrome, bandage, etc. Wishes, needs, etc.*—(1) Expressed wish, (2) unexpressed wish, and (3) precarious situation. *Adult in distress or want*—chiefly physical danger or physical need. *Animal in distress or want*—physical danger, need such as hunger, real or supposed attack by another animal or human being.

The author makes many summaries of social behavior, some in

[1] Lois Barclay Murphy, *Social Behavior and Child Personality: an Exploratory Study of Some Roots of Sympathy* (New York: Columbia University Press, 1937).

the form of verbal descriptions and some in the form of quantitative data. Table 9, illustrating nonvariate data, is given below.

The techniques and problems suggested by these studies are similar to those of other nonvariate qualitative and quantitative studies reported in the literature. When interviews, inventories, and questionnaires are used, the categories are in a sense already established by the instruments themselves. This fact however does not relieve the investigator of the responsibility of establishing suitable categories.

TABLE 9. Sympathetic Responses Occurring in 216 Hours of Records of Language (Records from a study by M. S. Fisher)

Form of Behavior	*Frequency in Records of 216 Hours of Behavior*	*No. of Children Who Showed This Behavior*
Helps child in need of assistance (not distressed)	21	12
Helps distressed child	8	7
Asks crying child why he cries	21	16
Sympathetic comment	20	14
Asks teacher why child cries	17	9
Looks at crying child	14	12
Warns child	16	8
Verbal defense of child who is attacked	11	8
Active defense of attacked child	15	8
Verbal comfort	2	2
Active comfort	6	5
Sympathy for toys or pictures	8	4
Miscellaneous	10	9

Graphical methods for presenting nonvariate data. Nonvariate data frequently can be shown to advantage by the use of graphical methods. There are many graphical devices available: bar, graphs, pie graphs, belt graphs, pictorial charts, etc. It is not our purpose to discuss the techniques of graphical representation. Such discussions can be found in any one of a number of standard treatises on the subject.[1] If appraisal is the purpose of the investigation, it

[1] For additional materials on graphic methods, see the following:

American Society of Mechanical Engineers, *Engineering and Scientific Graphs for Publication* (New York: American Society of Mechanical Engineers, 1943).

W. C. Brinton, *Graphic Presentation* (New York: Brinton Associates, 1939).

Bruce L. Jenkinson, *Bureau of Census Manual of Tabular Presentation* (U. S. Government Printing Office, Washington, D. C., 1949).

F. S. C. Northrop, *Logic of the Sciences and the Humanities* (New York: The Macmillan Company, 1947).

Rudolf Modley, *How to Use Pictorial Statistics* (New York: Harper and Brothers, 1937).

will be necessary to show not merely the status data but the standards with which comparisons are made.

Mathematical descriptions and appraisals of status employing variate data. Our concern in this section is with status studies employing variate data. Such data may be secured from applications of various rating scales and many mechanical measuring devices and tests. The problems of principal concern are those of (1) defining the purposes of such surveys; (2) choosing appropriate data-gathering devices; (3) collecting the necessary data; (4) tabulating and summarizing data, and (5) appraising results.

Defining the problem. Having defined an area (aspect of the educational program) that one desires to investigate, one must next fix clearly in mind the specific questions to be answered or hypotheses to be tested. Suppose, for example, that one has chosen to investigate the motivational factors that operate upon sixth-grade pupils in the one-room rural schools of Clay Township. One of the investigator's first tasks is to come to grips with the term "motivational factors." What is a factor? What is a motivational factor? Presumably distinctions will be made between incentives and motives; between purposes and motives; between values and motives; and between attitudes and motives. Some delimitation of the problem may be necessary. Presumably we are interested in factors whatever their source: home, school, or community. As the problem is stated, it would appéar that we are interested in motives of all types: good and bad; specific and general; immediate and remote; tangible and intangible. The research worker can place whatever restrictions upon his problem that he may desire. In time the general statement of the problem might be analyzed into a series of specific questions to be answered: Are sixth-grade children attending one-room rural schools in Clay Township more frequently motivated by concrete objects or by abstract values? What kinds of motive appear to arise out of home background? Out of school background? Out of community background? Are there differences in the motives of boys and girls? Of the children of tenants and of landowners? Of parents who attend church and of nonchurch goers? Are there differences in the motives of low and high achievers?

To answer all of the questions stated above would take the investigator well beyond the ordinary status study and into many complex patterns of interrelationship such as those to be studied in later chapters of this book. We are almost always interested in these complex interrelationships even though we attempt for

the time being a simple, close at hand, immediate status study.

Choosing appropriate data-gathering devices. The data-gathering devices that one may employ in variate status studies are numerous. The major characteristics of these devices have already been discussed in some detail in earlier chapters of this book.

All data-gathering devices must be validated not merely in general but for the immediate situation. The techniques of validation vary from one data-gathering device to another and for the various types of information sought. In general, however, one performs a completely new validation or as much of one as seems necessary for the immediate purpose.

Collecting the necessary data. There are many general and special conditions that should be held constant in applying data-gathering devices. In a ball-tossing experiment, for example, we try to keep certain conditions constant. It would be unusual to permit the participants to proceed under very different conditions: some singing, some counting, some walking about, some sitting, some with spectators, some without, and so on through the many other conditions that may affect test results. Time may be an important factor to hold constant. But there may be many other equally important conditions inherent in the situation or its relation to the data-gathering devices that should be held constant if meaningful results are to be obtained.

Tabulating and summarizing data. The statistical techniques employed in the analysis of status data commonly consist of the various measures of central tendencies and variability. Procedures for calculating these values are provided in most elementary books on statistics.[1] An example of test data taken from a recent study of the performance of college preparatory pupils in public secondary schools is given in Table 10. This is only one of a series of tables providing comparative data for a number of public and independent secondary schools.

Table 10 shows that certain values have been calculated: the range, first quartile, the median and the third quartile. In Table 11, the author reports the mean score, standard deviation, the standard error of the mean, the differences in the means, and the critical ratio. These are the statistical values; (that is, those

[1] Readers of these materials will vary greatly in the amount of statistical training and experiences that they may have had. For the completely uninitiated statistically, *An Educational Statistical Primer,* by Carlson and others (Dembar Publications, Inc., Madison, Wisconsin) is recommended. A more advanced, but still very readable treatment of elementary statistics will be found in Helen Walker's *Elementary Statistical Methods* (Henry Holt and Company, New York).

TABLE 10. Distributions of Scores of Grade XI College-Preparatory Students on Co-operative English—Test A: Form Y in Ten Public Schools[1]

September, 1949

Score	*I*	*H*	*R*	*L*	*P*	*K*	*M*	*O*	*N*	*Q*
96										
94										
92										
90				1						
88										
86				1						
84										
82										
80										
78					1					
76	1			2	1	1		1		
74	2			1	1			2		
72	1		2	1	3			2		
70	5	2	1	1		2			2	1
68	3	1	1	2	2	2		2	3	1
66	11	3	5	4	4	7	3	3	3	1
64	16	2	9	6	7	1	1	3	1	2
62	19	4	9	7	8	8	2	4	1	3
60	22	5	11	8	13	8	5	9	4	3
58	25	6	12	6	22	12	3	12	10	5
56	26	6	11	9	18	9	4	12	7	7
54	22	3	10	9	14	13	6	4	10	7
52	19	5	4	6	20	9	5	3	14	6
50	12	2	7	15	21	6	5	7	13	14
48	19	2	7	8	11	10	6	14	19	5
46	6	3	11	9	18	11	2	7	13	8
44	13	3	6	8	18	4	6	16	20	9
42	3	3	6	6	12	11	5	13	14	12
40	2	1	3	2	6	6	2	6	10	16
38	1	1	5	5	12	14	2	6	6	10
36	1		5	2		2	2	6	17	11
34			3	2	3	5	2	7	6	8
32				1	2	2	1	4	4	7
30	1					2	1	2	3	6
28						3			2	3
26	2							2		3
24										
22										
20										
18										
16										
14										
Total	232	52	128	122	217	148	63	147	182	148
Q3	62.0	61.6	61.1	60.9	58.7	58.7	57.1	58.2	53.4	51.9
Md	57.2	57.0	55.4	53.0	52.6	51.3	51.0	48.6	47.4	43.7
Q1	51.7	50.0	46.7	47.0	46.1	42.5	44.3	42.6	41.5	37.8
Range	26–77	38–71	35–73	32–90	33–79	29–77	30–67	27–76	29–71	26–70

[1] Taken from Robert Jacobs, "A Study of the Need for Special Norms on Scholastic Aptitude and Mechanics of English Tests for College Preparatory Students in Public Schools." In the *1949 Fall Testing Program in Independent Schools and Supplementary Studies* Educational Records Bulletins, No. 53, New York: Educational Records Bureau, 1950. pp. 52–66.

TABLE 11. Significance of the Differences between Mean Scores Obtained by Independent and Public-School College-Preparatory Pupils in Grades X and XI on the American Council Psychological Examination, 1948 Edition and on the Co-operative English Test A: Mechanics of Expression, Form Y

Test	*Group*	*Grade*	*N*	*Mean Score*	*S. D.*	*S. E. Mean*	*Diff. in Means*	*S. E. Diff.*	*C. R.*
A. C. P. Exam	Ind.	10	3412	100.62	21.31	.365			
Total Score	Public		687	97.17	18.99	.724	3.45	.810	4.26
	Ind.	11	4067	110.87	22.12	.347			
	Public		1537	102.15	22.92	.584	8.72	.680	12.82
L-Score	Ind.	10	3416	59.75	13.90	.238			
	Public		688	55.98	12.00	.457	3.77	.516	7.30
	Ind.	11	4069	66.61	15.18	.238			
	Public		1537	60.54	15.19	.387	6.07	.455	13.34
Q-Score	Ind.	10	3414	41.35	10.19	.174			
	Public		687	41.74	9.79	.374	−.39 *	.412	.96
	Ind.	11	4068	44.74	9.87	.155			
	Public		1538	42.13	10.60	.270	2.61	.312	8.38
English Test A.	Ind.	10	2365	52.81	8.62	.177			
	Public		444	47.14	8.37	.397	5.67	.435	13.04
	Ind	11	2498	56.53	8.47	.169			
	Public		1439	51.68	9.94	.262	4.85	.312	15.55

* Difference in favor of public-school group.

shown in the two tables) ordinarily calculated in such studies. The author concluded that this special group was somewhat differentiated from the independent school population, more so with respect to verbal ability and mechanics of expression than skills such as those measured by the quantitative sections of the American Council Psychological Examination.

Appraising results. After data have been collected and organized, judgments are made in terms of norms, criteria, or individual conceptions of acceptability. These standards of reference are just as important in making valid value judgments as the data themselves. Seashore and Ricks discuss at some length the importance of appropriate norms.[1] They make the following points:

1. Norms should yield meaning in terms of the particular purposes for which the testing is done. It may not be meaningful to teach for one set of purposes and test for another. In any case the conditions under which a given set of results are achieved should be noted.
2. Avoid unjustified general norms. General national norms may or may not be appropriate or helpful. People-in-general norms are legitimate only if they are based upon careful field studies with appropriate

[1] Harold G. Seashore and James H. Ricks, Jr., "Norms must be relevant," *Test Service Bulletin,* The Psychological Corporation, No. 39, May, 1950.

controls of regional, socio-economic, educational, and other factors in achievement.

3. Define national norms in terms of subgroups when possible. So-called national norms should be expressed in terms of sense making subgroups with the sampling procedures described in sufficient detail for understanding and evaluation by others.

4. Combine populations with care, and only when the resulting group has definite meaning. Avoid combining incongruous data into nonmeaningful, ambiguous larger groups. Small and ill-defined groups should be omitted.

5. Report all the genuinely useful data available. An enormous variety of educational occupation and clinical group data are needed to provide all the possible and desirable sets of norms.

6. Accumulate and use local and special group norms. Published norms may be helpful but they should be supplemented by such local and special data as are available.

The suggestions above pertain to test norms. Norms are not, however, always expressed in quantitative terms. Besides the nonquantitative individual standards of acceptability carried in the minds of most people there are many carefully defined behaviorial norms. Such norms have been widely used in clinical psychology and personnel work. Barr [1] and others have attempted to describe the good teacher by listing the specific behavior to be expected of him. Charters [2] attempted to define character traits in terms of the specific behavior to be associated with different purposes and situations. A committee from the several regional accrediting associations have developed the widely used *Evaluative Criteria.*[3] Similar criteria have been developed for the evaluation of guidance activities.[4] Shane [5] has compiled in this area a bibliography of current investigations which the reader should consult for further illustrative material.

A summary statement of the requirements for good mathematical descriptions and appraisals of status. At the close of the preceding section of this chapter a summary of the requirements for developing good verbal descriptions and appraisals of status was

[1] A. S. Barr, William H. Burton and Leo J. Brueckner, *Supervision* (New York: D. Appleton-Century Company, 1947).

[2] W. W. Charters, *The Teaching of Ideals* (Macmillan: New York, 1927).

[3] *Evaluative Criteria*, Co-operative Study of Secondary School Standards, 1950 Edition (744 Jackson Place, Washington, D. C., 1950).

[4] *Criteria for Evaluating Guidance Programs in Secondary Schools*, Occupational Information and Guidance Service, Division of Vocational Education, Office of Education, Federal Security Agency (Washington, D. C.: 1949).

[5] Harold G. Shane and Seymour Rovner, *A Selected Annotated Bibliography of Evaluation Instruments and Related Materials* (School of Education, Northwestern University, Evanston, Illinois, 1950).

provided. While mathematical symbols are used in the development of quantitative descriptions of status, to supplement verbal symbols, no essentially new logical processes are introduced beyond those already summarized at the end of the previous section. The reader might well re-examine the list provided there in light of the additional content provided in this section and the illustrations employing mathematical treatment of data. Particular emphasis has been placed in this section upon the establishment of categories and the appraisal or interpretation of data.

APPROACHES TO THE APPRAISAL OF STATUS

Appraisals may be made from many points of view. It has already been pointed out that status studies may be made of products, processes and conditions. It has also been emphasized that status is evaluated through comparisons with criteria, norms, or other objects of the same category. Within this frame of reference there are other considerations that need to be kept in mind as status is evaluated. In an earlier chapter emphasis was placed upon the fact that evaluations should be made with reference to educational needs and purposes. This is very important but status may also be evaluated with reference to persons and conditions. With three aspects of the program, namely products, processes and conditions to be evaluated from many points of view the process is exceedingly complex.

Products may be variously viewed. It is not uncommon to have the educational outcomes achieved by some school system, school, or class of pupils evaluated from the point of view of purposes other than those that guided the learning experiences of those evaluated. Notwithstanding the wide use of the *Evaluative Criteria*, which emphasizes the fact that evaluations must be made from the point of view of the stated objectives of the teachers, schools, or school systems being evaluated, many persons today naively or unwittingly continue to appraise educational outcomes by inappropriate standards. Products may also be considered with reference to persons. No one expects the same kinds and amounts of growth and achievement from very different kinds of persons. We shall have more to say about this in another chapter when we discuss case studies. Much the same can be said for conditions. We now have norms for noncollege preparatory pupils as well as for college-preparatory pupils; for rural and urban pupils; for pupils from private schools and pupils from public schools, for different states, races, and economic groups.

The process may also be a consideration in appraising products. Such questions as the following may be asked: Is the product the result of pupil resourcefulness or of teacher direction? Is the product the result of teacher centered and dominated activity or of pupil centered and self-directed activity? Is the product the result of group activity or of individual activity? The methods by which products are produced are frequently of concern to those who attempt to assess the adequacy of status.

Processes may also be appraised from different points of view. Some processes are more in accord with certain purposes, persons, and conditions than others. If one is concerned with democratic living as a goal of school education, then democratic processes are appropriate. Processes, methods, and procedures must also be appropriate to the individual. Patterns of behavior, be they social, physical, or intellectual, must constitute the best possible utilization of the assets and liabilities of each individual. Finally, processes always take place in a socio-physical setting. In many instances the acceptability of a process is a matter of custom or tradition or of social practice. Rightly or wrongly processes ordinarily have social foundations. To view things merely as efficient means to immediate ends is not enough. This point will be further developed in the discussion of social foundations in a subsequent chapter. The physical environment imposes other restrictions.

Conditions may be appraised from different points of view. First we are concerned with conditions that limit or facilitate educational outcomes; whether a condition is important depends upon the goal sought. Beyond this the adequacy of a particular condition is largely an individual matter. There are both likenesses and differences among individuals. The likenesses give rise to the principles of science. Insofar as all individuals are alike, personal factors may or may not influence the appraisal of conditions, but to the extent that each individual is unique, personal factors may become exceedingly important in appraising conditions thought to influence pupil growth and achievement. The adequacy of a given condition may be a function of the unique characteristics of the individual considered. Finally, conditions are hierarchial in character. For every condition, however immediate or remote, there are other more remote or fundamental considerations. Back of these are considerations that constitute important points of view from which the more immediate aspects of life need to be evaluated.

The reader should be careful not to lose the connection between this and the succeeding chapter on sampling surveys. Whether a survey is of the sampling or nonsampling type, one is involved in

the description and ultimately in the appraisal of status. The sampling survey provides a means of studying certain types of problems where one's concern extends beyond the immediate groups at hand.

SUMMARY

Status studies may be made of products, processes, or conditions. To conduct a status study, there must be appropriate instrumentation, adequately defined goals, and a frame of reference. The ultimate goal of the school program is pupil growth and achievement. This provides one point of view from which appraisals may be made. Adequacy is also a function of the person and the setting. These two constitute important points of view from which appraisals may be made. Finally appraisals are made through the use of criteria that may not immediately or directly involve any of the above. The appraisal of status may be a very complex activity involving many interrelationships.

CHAPTER VI

The Sampling Survey

The method of investigation by sample has for its purpose the description of the properties of an accurately defined population by means of the information obtained from the sample. Sampling, that is, the selection of a part to represent the whole of a population, is a procedure of long standing and importance. It is indeed the most important problem in practical research. If there were no validity in the use of samples in scientific inquiry, investigations would be impossible unless the total population were studied. If this were a necessary condition, it would preclude most investigations. Even the Federal government with its resources seems to find it practicable to obtain a complete census only every ten years. Even when a complete inventory becomes possible it is probably more exact to speak of the population as a 100 per cent sample since the findings are generally used for the purpose of drawing inferences about situations presumably comparable to those under which the data were originally collected.

There are at least three major reasons for the very rapid development in the use of samples in obtaining information:

1) *Reduced costs.* Expenditures are obviously smaller when data are obtained for only a small part rather than for the whole of a population.

2) *Greater speed.* Data can be more rapidly collected, processed, and published with a sample than with a complete enumeration of the population. This is often of vital importance when information is urgently needed. Witness, for example, the length of time before the findings of the Federal Census in published form or of Federal and state publications dealing with school statistics are available.

3) *Greater accuracy.* A sample may actually provide more accurate information than that provided by the kind of complete

study of a population that would prove practical at any given time. With changeable characteristics, speed in reporting may be essential to accurate up-to-date information. It may provide a more accurate account than a report of the entire population published much later; the latter in fact may have only historical significance.

Although sampling is a fundamental problem in many types of research, our primary purpose is to consider sampling in relation to what we have called the *sampling survey*. More particularly our concern will be with the design of the survey, which may be either descriptive or analytical.[1]

There is considerable confusion in literature regarding differences between an experiment and a survey. Both entail models by which observations are taken. In experimental designs the objective is to estimate the effects of differential treatments; in survey studies the purpose is to estimate certain characteristics for a specific population. The survey may be entirely enumerative in character or it may be analytic. It is analytic when it aims to establish the existence of associations in the population or when the interest is in the factors that may have been operative in producing an observed situation.[2] Only by means of an experiment can we establish with the certainty possible in science the magnitude in the causal sense of any given factor.

Although surveys cannot be considered as adequate substitutes for experiments, they are especially useful in situations in which it is very difficult or perhaps impossible to conduct an experiment. Surveys are particularly valuable in exploratory work preliminary to experimentation in that they may serve to identify factors that are worthy of experimentation. To serve these purposes, appropriate survey designs and statistical analyses must be applied.

The close relationship between experimental and survey designs is due chiefly to the fact that the same principles underlying experimental designs serve as the bases for the development of modern sampling designs. The central problem in the sampling survey is that of obtaining unbiased estimates of the quantities under survey and of measuring the errors of such estimates. Only the princi-

[1] Since modern survey designs and the corresponding methods of analysis require a functional knowledge of the principles of design and analysis, including the analysis of variance and covariance, the investigator proposing to use sampling surveys may wish to read more detailed discussion than we can present here. See especially:

William E. Deming, *Some Theory of Sampling* (New York: John Wiley & Sons, 1950).

Palmer O. Johnson, *Statistical Methods in Research*, Chapter IX (New York: Prentice-Hall Inc., 1949).

Frank Yates, *Sampling Methods for Censuses and Surveys* (London: Charles Griffin and Company, Limited, 1949).

[2] In this connection, compare the discussion on status studies in Chapter V.

ples of modern experimental design, particularly those of *randomization* and of *replication* together with the technique of analysis of variance made such development possible.

Early investigators were led to an appreciation of the need for estimating sampling errors from the results of their observations, both to determine whether the sampling method used was adequate for the purpose and to increase the efficiency of future sampling of the same kind of material. There were two conditions upon which their work was based:

(1) If the sample is to be unbiased, the units of the sample must be selected by some process which is independent of the characteristics of the individuals sampled; the observer must exercise no control in the choice of the elements of the sample.

(2) In order for an estimate of sampling error to be made available a minimum of at least two sampling units must be obtained from the material being sampled. Also, these sampling units must be selected at random from the complete aggregate of sampling units that comprise the bulk of material. The sampling units must be of approximately the same size and pattern.

The necessity for the first of these conditions had been recognized for a long time, but it was the methodological requirement for fulfilling the second that led to the development of modern designs. The process of *randomization* introduced into experimental design by Fisher furnished a valid estimate of sampling error. The technique of analysis of variance made possible the pooling of estimates of error and the separation of heterogeneous components of error. This provision made possible the reduction to a small number of the sampling units taken from the sampled material. It thus made possible the development of rather complex sampling designs which involve samples in two or more stages.

In the earlier application of modern principles, the nature of the material sampled was such that it could be subdivided into sampling units which were of approximately uniform size and shape. Accordingly, the sampling designs were relatively simple. Later, applications were made to problems where there were large differences in variability in different parts of the population and in which there were sampling units widely differing in size. Modern designs are capable of dealing with such situations.

REQUISITES OF A GOOD SAMPLE

The primary purpose of any sampling procedure is to obtain a sample which, within the restrictions imposed by its size, will re-

produce the characteristics of the population with the greatest possible accuracy. Accordingly, it might be thought that a deliberate selection of the sampling elements would yield the most accurate results. The state superintendent of schools might, for example, be asked by an investigator to select a sample of "typical" schools, which he might use to study the school population. Or, as is sometimes the case, he might specify "representative" schools which a foreign observer might visit.

Such samples, however, are of little value to a critical investigator. Their principal defect is that they are likely to be *biased;* i.e., the selection of the schools may have been influenced by similar errors. In order to enhance the reputation of a state or of a county, school authorities may tend to select all schools which are better than the "average." Even if school officials try to be entirely objective, unconscious errors of judgment all acting in the same direction may occur, outweighing any improvement in accuracy which might result from such deliberate selections. Neither could it be assumed that if a greater number of school officers should participate in the selection improved accuracy would result since all might be prone to error of the same type.

Thus we can distinguish between two types of sampling error: (1) those that result from biases in selection, and (2) those that are attributable to chance differences between the elements of the population which are included and excluded in the sample. The aggregate of the former type constitutes what is called *error due to bias,* and of the latter type *random sampling error.* The total sampling error is, therefore, comprised of errors of bias, if such exist, and the random sampling error. Although bias forms a constant component of error which does not decrease as the size of the sample increases, random sampling error decreases with sample size. The amount of the error depends upon design of the sample of which the size of the sample is one factor.

Since no objective conclusions can be drawn from samples that are biased it is necessary to obtain (as far as it is possible to do so) unbiased samples. It is important to know how such samples may be selected to avoid bias, and the methods of selecting samples which give rise to bias. We shall illustrate a number of procedures where faulty selection of the sample may introduce bias. The principal faulty methods are:

1) Deliberate selection of the units of the sample purported to be "representative." This type of bias has been discussed above.

2) Selection by a procedure where there is a connection between the method of selection and the characteristic(s) under con-

sideration: for example, selecting the inhabitants of a city from the names in a telephone directory if the variate of interest is the number of children attending college; or the selection of a sample from an alumni directory to study the occupational destination of graduates. Thus, some inhabitants, most often the poorer, may not have telephones. Less successful graduates may not mention occupations, or even be listed where listing involves subscription for the directory.

3) **Choice of a random sample.** Human bias is very prevalent. Even when aware of their own imperfections, trained observers may be biased; even in similar situations or circumstances, different observers may be biased in different ways. The same observer may display bias in different ways under different circumstances. Requesting teachers to select three random samples of their pupils to engage in a nutrition experiment wherein the home ration is to be supplemented by (*a*) pasteurized milk, (*b*) raw milk, or (*c*) no milk, may, for example, lead to bias.

4) **Substitution.** Sometimes investigators substitute one convenient sampling unit for another when there is difficulty in obtaining the desired information from the original selectees. In a house-to-house canvass, for example, the neighboring house may be substituted for the one at which nobody is at home. This practice would necessarily result in a disproportionate number of houses that are occupied throughout the day, e.g., homes of people with families.

5) **Incomplete coverage of the units selected for study.** If no follow-up is made to houses where there was no reply in previous visits, bias will be introduced even if no substitution is attempted as indicated in item 4. For example, one of the writers recalls reading a proposed master's thesis in which the student used the sampling survey and the interview technique. He was struck by the abnormal size of families in this town where he had once lived. Upon inquiry it was found that the student had summarized results after only one call. Obviously, the larger the family, the greater the probability that someone would be at home.

The defect is particularly prevalent in surveys where the questionnaire is used to collect information. In such cases respondents are likely to be those to whom the subject of inquiry is of special interest, or who possess other characteristics that make them peculiar in some respect. A good example of the effect of selection is a study reported by Reid,[1] who surveyed public school principals in

[1] Seerly Reid, "Respondents and non-respondents to mail questionnaires," *Educational Research Bulletin* (Ohio State University, Vol. 21, pp. 87–96, 1942).

Ohio with respect to use of the radio in their schools. From the initial mailing, 42 per cent of the principals responded. Reid made successive follow-ups until 95 per cent of the principals had replied. He found that the actual proportion of schools that owned and used radio equipment was greatly exaggerated in the early returns. Even when 65 per cent of the principals had responded, there was a marked bias in the basic estimates desired.

PROCEDURES IN SELECTING SAMPLES

Method of selection. The simplest way to avoid bias in selecting the elements of a sample is to draw the elements either entirely at *random,* or at random subject to restrictions that will improve the accuracy but not introduce bias into the results. Certain forms of *systematic* selection, such as the selection of names at uniform intervals down a roster may be satisfactory. When we employ *unrestricted random sampling,* the method is such that each individual in the population has an equal chance to be included in the sample. When we employ the method of sampling called *stratified random sampling,* the population is first subdivided into a finite number of strata. From each stratum we select a predetermined number of observations by random sampling. The method called *purposive selection* uses the principle of selecting individuals to be included in the sample according to some criterion or criteria called controls. This method was used predominantly at one time in sample surveys. But because of lack of rigorous rules of selection, which resulted in samples found to be by no means equivalent to balanced random samples and frequently unrepresentative in a number of ways, the method has been largely replaced by more thorough application of the principles of stratification, balancing, etc.

The research worker should always describe specifically the method of selecting his sample, since the words "random" and "random sample" are often gravely misused. The most reliable method of drawing random samples is by the use of random sampling numbers. There are currently three Tables of Random Sampling Numbers available: Tippett's, Kendall and Babington Smith's, and Fisher and Yates'.[1] In systematic sampling from lists, it is necessary to make sure that the lists are complete, accurate, and recent. Here the practice is to take, say, every *k-th* entry on the list. The first entry should be determined by selecting a number at

[1] Palmer O. Johnson, *Statistical Methods in Research,* Chapter IX (New York: Prentice-Hall Inc., 1949).

random between I and k. This aspect of randomness does not, however, convert the sample into a random sample. The arrangement of lists is not a random one. Perhaps the closest approximation to a random order is that given by alphabetical lists, though these may possess certain nonrandom characters, such as nationality and blood relationship.

Quasi-random samples, however, may automatically result in some sort of stratification, such, for example, as an electoral list arranged by wards and streets. In general, systematic sampling from valid lists may prove to be satisfactory if caution is taken to note that there are no periodic features in the list that may be associated with the particular sampling interval used. Systematic sampling from lists will usually give unbiased estimates of arithmetic means so long as the starting point is chosen at random. But no accurate estimates of the standard errors can be obtained from individual samples selected in this manner. Some recent research indicates that for certain special kinds of systematic samples standard errors may be estimated. If repeated systematic samples are taken from the same population, the observed variation in the means may be used to estimate the precision of such samples. The minimum number of samples should be two.

Bias in estimation. We have discussed biases which result from faulty methods of selection and from faulty procedures during the collection of the data. We should point out that faulty statistical methods of analyzing the findings may also introduce bias. There are three different criteria that must be taken into account in determining the best estimate to be had from any given type of sampling: (1) absence of bias, (2) accuracy or efficiency, and (3) computational convenience. If the population values from which the random sample has been chosen are normally distributed, the arithmetic mean will provide an unbiased estimate of greatest accuracy. Likewise, the unbiased standard deviation is the sufficient estimate of variability.

The problems, however, of bias and relative efficiency connected with the most useful estimates, that may arise from any given type of sampling call for the application of more advanced mathematical statistical theory, than can be presented here.[1] It should be

[1] See, for example, the following:

William E. Deming, *Some Theory of Sampling* (New York: John Wiley & Sons, 1950).

Palmer O. Johnson, *Statistical Methods in Research*, Chapter IX (New York: Prentice-Hall, Inc., 1949).

Frank Yates, *Sampling Methods for Censuses and Surveys* (London: Charles Griffin and Company, Limited, 1949).

noted that there are cases where a certain amount of bias, provided that it may be shown as reasonably constant, may be accepted, as, for example, in the comparison of different groups of a population where the bias is approximately constant from group to group. Also, there may be minor sources of bias, which can be tolerated if they introduce errors which are relatively unimportant in comparison with the kinds of bias discussed here and with random sampling error.

The control of random sampling error. Whether the sample may serve to define accurately the properties of the population will depend chiefly upon the amount of sampling error introduced by the sampling process. Even if the procedure of selection follows the canons of the random sampling process, the sample cannot be exactly representative of the whole population. The inevitable errors resulting from the process are called *random sampling errors.* The average size of these random sampling errors depends upon the size of the sample, the variability of the individuals sampled, the sampling design adopted, and the method of calculating the results. These sources of variation suggest ways of reducing the magnitude of the sampling error.

Aside from errors due to bias, the simplest means of increasing the accuracy of the sample is to increase the size of the sample. Other factors being equal, the size of the random sampling error is approximately inversely proportional to the square root of the number of units comprising the sample. The accuracy also depends upon the variability per unit of sampling; or more accurately, on that portion of the variability per unit that contributes to the sampling error. Modern sampling designs, while placing restrictions on fully random selection, serve to reduce the variability per unit contributing to sampling error and thereby to decrease the size of the sample required for a given accuracy.

The simplest kind of restriction is that of stratification. There are a number of other devices that one may employ to increase the accuracy of the sampling procedure. Three of the most important devices are (1) use of supplementary information, (2) use of a variable sampling fraction, and (3) multistage sampling.

Supplementary information involves the use of information obtained from sources outside the sampling scheme or from a more extensive sample than that on which information on the main characteristics is based. An example is the use of data on the occupations of parents and the occupational distribution of residents of the area served by a municipal junior college to determine the occupational representativeness of a sample of students entering a

municipal junior college. The use of a *variable sampling fraction* involves the inclusion of different proportions of the several strata in the sample, thereby making it possible to sample more intensively the more important, or more variable, parts of the population.[1] In *multistage sampling*, the population is first classified into a number of first-stage sampling units, sampled in the usual manner. The selected first-stage units are then subdivided into smaller second-stage units and sampled. Further stages may also be added if desired.[2] Thus, for example, in a school survey, a sample of cities might be taken. For each of the selected cities a subsample of schools might be taken, with, possibly, a further subsample of classes from the selected schools.

Type and size of sampling units. Sometimes the population can be variously classified into units. Thus we might consider a city as composed either of a number of city blocks, or of a number of households, or of a number of persons. In general, when a given proportion of the population is included in the sample, the smaller the sampling units employed, the more accurate and representative will be the sample results. For example, in a state school survey, it will be more accurate to take 20 per cent of all schools in each county, than to take all the schools in 20 per cent of the counties.

A change in the type of sampling unit will usually affect both the cost of taking and the accuracy of the sample. The best unit is the one which gives the desired variance for the sample estimate at the least cost. Particularly where the interview is to be used, the need for small units distributed over the whole of the population is frequently in conflict with administrative requirements. It is obviously easier to arrange for a survey of schools in compact areas, the county, for instance, than to survey the same number of schools scattered over an entire state. To obtain a satisfactory balance between these two conflicting requirements is frequently one of the central problems in the planning of a sample survey. Consequently, sampling designs which might be excellent for questionnaires might be undesirable when the information is collected by special investigators.

The term *cluster sampling* is often applied to sampling in which the sampling units are aggregates or "clusters" of the natural units. The smallest practical or feasible units for certain educational investigations is the school class. With such units special problems

[1] Francis G. Cornell, "Sample plan for a survey of higher education enrollment," *Journal of Experimental Education*, 1947, 15:213–218.

[2] Frank Yates, *Sampling Methods for Censuses and Surveys* (London: Charles Griffin and Company, Limited, 1949).

arise in obtaining statistical estimates and the measures of their sampling errors. For example, cluster sampling almost always increases sampling error as compared with unrestricted sampling error of the same number of cases. This is due to the sampling of previously existing groups (classes, for instance) of the population which involves a positive intraclass correlation of the variable under investigation.[1]

PLANNING A SAMPLING SURVEY

There are many practical problems encountered in sampling surveys, such as those met in the study of certain educational problems. These problems have been rather arbitrarily grouped under ten categories, which are, in general, discussed in the order in which they are confronted in a practical situation. The classes of problems cannot be considered as independent because any decision made in a given case will likely influence more or less the decision taken with respect to others. They, therefore, need to be considered jointly; where independent judgments are formulated they should be considered as tentative until the whole plan has been finally formulated.

The ten classes of problems are as follows:

1) Statements of the objectives of the survey
2) Definition of the population or populations to be sampled
3) Determination of the nature of the data to be collected
4) Techniques of collecting the data
5) Selection of frame and sampling unit
6) Method of selecting the sample
7) Treatment of the nonrespondents
8) Conducting the pilot or exploratory surveys
9) Summary and analysis of the data
10) Preparation of the sampling survey report

1) Objectives of the survey. The purpose of collecting data is to afford a basis for action. All questions that have meaning are raised for sake of some purpose. Problem solving is an aspect of purposive planning.

Careful consideration, therefore, should be given to the purposes for which the survey is to be undertaken and the uses to be made of the findings. Are the results required primarily for administrative or research purposes? When the object of the survey is an

[1] Eli S. Marks, "Sampling in the revision of the Stanford-Binet Scale," *Psychological Bulletin*, 1947, 44:413–434.

administrative decision only, there is not the question of impersonal validity as is the case where the aim of the investigator is the advancement of scientific knowledge. The administrator who is responsible for the decision plans the best survey possible within the time available. He then makes a decision which he thinks best. He is not interested as to whether the method of inquiry is certain to lead to the correct decision in the long run.

A clear idea is needed at the outset concerning what is to be found out. The task is to ascertain as accurately as possible what information is required for the purpose. There is needed a method of action which will enable the investigator to collect the observations pertinent to the questions asked.

2) Definition of the population or populations to be sampled. The investigation may be concerned with the estimation of characteristics of a single population or with comparison of analogous characteristics of two or more populations. This is determined by the purpose of the study. A necessary second step is a clear definition of the population or populations about which it is desired to draw conclusions. If this is not done, the sample studied often may be inappropriate.

In some cases, there may be no difficulty in defining the population, e.g., the graduates of a certain university. On the other hand, rules may be required to define what constitutes a "student," a "junior college," a "secondary school," or a "volume" in the library. The investigator should be able to decide without much hesitation whether a doubtful case belongs to a specified population. It may not always be possible to have the population sampled identical to the population about which information is sought. In predicting an election, for example, what is wanted is a random sample of the population of voters' opinions upon going to the polls. What one obtains is the population of opinions at sometime before election of how eligible voters intend to vote if they go to the polls. Both the opinions and the intentions of the sample of prospective voters are subject to change.

Definitions of the population must also be considered in conjunction with the selection of the *frame* (see item 5). The *frame* adopted has its own implied definitions of the types of materials to be covered. If the frame does not include certain classes of material such categories should either be omitted entirely or the frame supplemented.

3) Determination of the nature of the data to be collected. It is essential to have a clear idea of what we desire to find out about the population. What should we like to know? The data to be

collected depend upon the purpose of the inquiry—whether the results are to be used primarily for an immediate administrative decision or whether they are to be used for the advancement of scientific knowledge.

The principle is to plan so that the items on which information is sought form a rounded whole covering a specific subject or a logically consistent group of subjects. This principle is of particular significance where questionnaires are to be filled out by respondents or where the information is sought by field investigators. Valid and reliable information is obtained only if the respondents are able and willing to co-operate. There must be a clear purpose indicated; this purpose should be explained to the respondent; the questions should also be relevant to this purpose. If the sense of this importance is not realized and if the data are of a miscellaneous character, the respondents are not likely to give their best. Occasionally, when the questionnaires become unduly lengthy, it may be feasible to subdivide them and thus secure information on one subdivision for one group of respondents and on another subdivision for another group. Certain basic information will be asked of all the respondents and the two subdivisions can make up a part of interlocking samples. In this procedure, however, only the relationship between items of information in the two sets of respondents can be analyzed for certain strata but not for the individual respondents.

In arriving at a decision with respect to type of information required, collaboration of experts on the subjects under inquiry should be enlisted. In this way, besides the critical evaluation of the proposed items, one guards against the omission of what may be vital items of information. Likewise, the sampling plans and design of the investigation should be submitted for the critical evaluation of a statistician. A pilot study should be a part of the design. This problem will be discussed later.

Sometimes information is sought through direct observation or physical measurement. Here the points for consideration are whether the investigator or other individuals who will take or make the observations are competent and whether excessive amounts of time or excessively expensive apparatus will be required; also whether the owners of the surveyed material will permit measurements or other observations to be made. When the ideal cannot be achieved, there are at times possibilities for obtaining information which may correlate highly with the unattainable or unavailable quantities desired. The efficiency of such substitution, however, can be properly determined only by the appropriate

statistical investigation of the relation between the primary and secondary quantities.

4) **Techniques of collecting the information.** The devices used to collect the information are to a considerable extent conditioned by the nature of the material under investigation and the type of information sought. In general, observations are to be preferred to questions; questions of fact or those relating to past actions should have preference over those involving generalities or hypothetical future behavior. It is difficult to prescribe any general rule with respect to obtaining or making physical measurements and qualitative observations. The former are more objective but the latter are more effective in presenting the conspicuous points of a complex situation. Opinion in itself is meaningless unless one can predict from it what people actually will do. For example, what the respondent tells the interviewer—his overt opinion—may not necessarily be the same as what he really believes or will do—his covert opinion. Moreover, it seems that with some individuals it does more for one's ego to express an opinion—any opinion—than to acknowledge that one has no opinion at all. About three years ago, *Tide* magazine reported that an investigator, with tongue in cheek, carried out an opinion survey on the "Metallic Metals Act." There was no such act, but 70 per cent of those who were questioned expressed an opinion on it.

The data that are needed may be sought by house-to-house canvass or by other types of interviewing, by mailing or by reference to existing records of information. Sometimes several sources are combined in the same inquiry.

The mail questionnaire is frequently used in surveys because of the economies involved. The principal objection to this technique of collecting information is that it generally involves a large nonresponse rate and an unknown bias in any assumption that the respondents are representative of the combined total of respondents and nonrespondents. On the other hand, personal interviews generally yield a substantially complete response but at a cost per schedule, which is considerably higher than that for the mail questionnaire.

Consideration of the requirements of the problem under investigation determines the technique to be used. Interview techniques may take various forms. One which has been found effective involves the following features: (1) specific questions are formulated so that all respondents are asked questions in a uniform manner but which can be answered by them in their own words, expressing shades of opinion and degrees of certainty or uncertainty,

and presenting reasons for the opinions and attitudes they possess; (2) the interviewer is trained so that he is able to conduct the interview in a conversational manner and to establish good rapport with the respondents; and (3) coding techniques are derived for the quantification of opinions expressed by the respondent in his own words. Techniques of analysis that afford objective checks of the survey data are developed.

When the sample questionnaire survey is used to collect information, special care is needed in framing the questions. This needs to be done at the planning stage, since the information elicited in the survey is dependent on the exact form of these questions. Likewise, where observation and physical measurements are to be used their exact form of collection should be determined during the planning stage.

If a question is to have meaning, it is necessary to be able to institute as definite a series of actions as possible whereby a set of relevant observations may be obtained. The construction of a questionnaire presupposes a formal characterization of respondents and their responses. Every effort must be taken to guard against a result where the "findings" are not simply the consequence of implicit assumptions suggested by the respondent when the investigator frames his questions. In questions of opinion, every effort should be made to formulate wording that is "neutral," that is, wording that does not bias the respondent to give one kind of answer rather than another. If a choice cannot be made between two different wordings, each wording may be used in half the questionnaires.

The order of the questions should receive careful consideration. The respondent will likely react more favorably to an orderly sequence. The investigator's task is also usually simplified by such an arrangement. In surveys, it is often desirable to give the respondent an opportunity to record general remarks on special points. These remarks serve to direct attention to relevant facts which the questionnaire does not include.

5) Selection of frame and sampling unit. Sampling units form the basis of the actual sampling procedure. Before the units can be unequivocally defined a frame must exist; or, if it does not exist, it must be constructed. For example, in the sampling of a human population in which the household is the unit of sampling, there must be available a list of all the households in the population to be studied. This list must make it possible to locate without ambiguity any household selected from it. In the sampling of schools, when the school is a unit, there must be available a complete list

of schools. The specification of the frame implies that the geographical scope of the survey is defined as well as the categories of material to be covered. If other categories of the population are required or if the frame is incomplete or otherwise defective, special means must be taken to supplement or correct the frame. If one were to sample all the rural schools of the United States, a complete list of all the schools would be required. Such a complete list is not now available. If such a sampling survey were to be inaugurated it would be necessary first to compile a complete list of rural schools.

The frame actually determines to a considerable extent the entire structure of the sampling survey. Consequently, until information is available concerning the nature and accuracy of the available frames, no detailed planning of the survey can be undertaken. If no frame is available, the construction of one appropriate for the purposes of the survey may well constitute a greater part of the work of the investigation.

An investigator should examine available frames from the standpoint of the following: Is the frame (1) inaccurate, (2) incomplete, (3) subject to duplication, (4) inadequate, or (5) out of date? Since many defects are not apparent until a detailed investigation has been made, the investigator should conduct a critical investigation of any frame which he plans to use. Such an investigation will involve a study of the administrative machinery to determine how the frame was constructed and how it is kept current. Such an investigation may also require some field work.

The size of the sampling unit is often influenced by the type of frame available or that can be made available for the survey investigation. Usually in educational surveys, lists of schools, teachers, principals, and school superintendents, and sometimes of classes and of pupils can be compiled from the records in the state department or of school systems. In a recent state-wide study, in which one of the writers was a consultant, a ninth-grade English class served as the most useful practical sampling unit; it was necessary to prepare a list of all such classes in the state. In a state-wide committee established to make recommendations concerning the preparation of secondary school teachers, one of the writers, as a member of the committee, was asked to prepare, send out, and analyze the returns from a sampling survey, within approximately one month. The data were collected and analyzed and presented at the next meeting of the committee. The study illustrates a number of points outlined in this discussion.

Only a postcard addressed to the individual teacher, principal,

or superintendent explaining the purpose of the study and an attached postcard self-addressed were sent out. After listing the number of years of experience and the position, each school official answered the following two questions:

I. Should the training of future teachers extend over (a) __4; (b) __5; (c) __more than 5 years of college work? Check one.
II. If you checked Ib or Ic, check one of the following: Additional work beyond the conventional four years of college should be distributed among 1) *teaching field,* 2) *professional education,* and 3) *general education* in which order of emphasis? Check one:

__a) 1–2–3	__d) 1–3–2	__g) equal distribution of 1, 2, 3
__b) 3–2–1	__e) 2–1–3	
__c) 2–3–1	__f) 3–1–2	

Three populations were sampled: (1) secondary school teachers, (2) principals, and (3) superintendents. A random sample was taken of each from the complete lists in the files of the State Department of Education. The order of the items of the second question was randomized. The returns were practically complete. The findings are presented in Table 12. This is an example of the use of survey results for administrative action. It shows how a sampling survey can be useful for such purposes. It would have been impossible, as well as unnecessary, to canvass the total population and make the results available within an interval of a month or less.

Other frames sometimes used, particularly for censuses and surveys of human populations include (1) lists of individuals in the population, or in subdivisions of it, provided for administrative purposes; (2) aggregates of census returns resulting from a complete census, e.g., the selection of 1 in 20 individuals for the collection of supplementary information collected on the spot by the census taker in accordance with certain well-defined rigorous rules in order to avoid bias; (3) lists of households or dwellings in given areas; (4) town plans; (5) maps of rural areas; (6) lists of towns, villages, and administrative areas, often with various types of supplementary information; and (7) master samples, e.g., the sample constructed by the Statistical Laboratory of Iowa State College, in co-operation with the Bureau of Agricultural Economics and the Bureau of the Census, comprised of 67,000 areas which identify about 300,000 farms located within nearly every county in the United States.

6) Method of selecting the sample. There are now available a variety of methods by which a sample may be selected. The method selected should provide the desired accuracy at minimum cost. The

size of the sample is important. The size of sample required for a predetermined precision can be estimated, at least roughly, when the method of sampling has been chosen and its sampling properties investigated. The determination of the size of sample required for a specified accuracy is relatively simple when a random sample is taken. The calculations are more complicated with the more complex methods of sampling which require more informa-

TABLE 12. The Preferences of a Random Sample ($N = 369$) of Schoolmen with Respect to Extended Preparation of Teachers

Years of Experience	*Total Numbers*	*4 years Training*	*5 years Training*	*More than 5 years*	*5 Years Training* * *A*	*B*	*C*	*D*	*E*	*F*	*G*
1 year	33	13	20	—	6	4	—	4	4	2	—
2 years	27	10	16	1	9	—	—	4	2	—	1
3 years	17	5	12	—	4	1	2	1	1	3	—
4 years	9	2	6	1	2	1	—	1	2	—	—
5 years	6	1	5	—	2	—	2	1	—	—	—
6–10 years	49	16	32	1	10	2	2	7	2	5	4
11–15 years	49	11	36	2	8	5	7	7	7	—	2
16–20 years	29	13	15	1	4	1	2	4	1	1	2
Over 20 years	76	25	47	4	9	4	2	13	7	8	4
Superintendents	34	9	25	—	2	4	3	5	6	4	1
Principals	36	5	31	—	7	4	7	4	5	3	1
No Report	4	3	1	—	—	—	—	—	—	1	—
Totals	369	113	246	10	63	26	27	51	37	27	15

* If the teacher checked five years of training as the desirable period of time to extend teacher training he was requested to check also in what order of emphasis additional work beyond the traditional four years of college should be distributed. For the "More than 5 year group" explanation is given under * note at the bottom of the page.

A—Teaching field, professional education, general education
B—General education, professional education, teaching field
C—Professional education, general education, teaching field
D—Teaching field, general education, professional education
E—Professional education, teaching field, general education
F—General education, teaching field, professional education
G—Equal distribution of teaching field, professional education, and general education

* NOTE: For the group of teachers who checked more than five years, the frequency of the order of training preferred was A-4, C-1, D-1, E-1, F-1, G-2.

tion of the population being sampled.[1] The methods of reducing the random sampling error, previously described (stratification, variable sampling fraction, and supplementary information) generally may be expected to increase accuracy. Accordingly, the estimate of the number of sampling units required where a random sample is to be used may be regarded as an *upper* limit to the number required with these other methods of sampling when the

[1] Marilyn Harris, D. G. Howitz, and A. M. Mood, "On the determination of sample sizes in designing experiments," *Journal of the American Statistical Association,* 1948, 43:391–402.

Palmer O. Johnson, *Statistical Methods in Research,* Chapter IX (New York: Prentice-Hall, Inc., 1949).

Frank Yates, *Sampling Methods for Censuses and Surveys* (London: Charles Griffin and Company, Limited, 1949).

same sampling unit is employed. The nature of the phenomenon observed may also have an effect on the size of sample. For example, a survey of the eye movements of a few readers may be sufficient, since there may be little variation from child to child with respect to the general characteristics of eye movements while reading. The relation of method to frame has been considered under item 5.

The purposes of stratification are two-fold: (1) to increase the accuracy of the over-all population estimates and (2) to secure suitable representation of the subdivisions of the population that are in themselves of interest. If a heterogeneous population is divided into homogeneous strata, the accuracy of the sample can be increased if there are marked differences between the different strata. Usually the increase in precision is greater for quantitative than for qualitative characteristics. The greatest over-all precision will be achieved if the strata are so determined that the sampling units within each stratum are as homogeneous as possible. If after the first subdivision there is still marked heterogeneity within certain strata, it is possible to stratify these into smaller more homogeneous subdivisions for the purposes of sampling.

In estimating the sampling error of a stratified sample, the variability between strata must be removed from the estimate of the variance of a single unit of sampling. This can be done by the analysis of variance technique.

If the population is classified in the required strata, the requisite number of sampling units from each stratum are chosen at random. A population may be stratified on the basis of two or more different characters. If selection of sampling units is made from substrata comprised of the various combinations of the main classification, the substrata are equivalent to strata, and the procedure of sampling is identical to ordinary stratification. The analysis of variance developed for use in multiple classification with unequal representation in the subclasses, can be used if the sample is stratified for two or more factors without control of substrata. That is, the elimination of the variability due to the separate factors is obtained by fitting constants for these factors. The approximate method developed by Tsao may also be used.[1]

We shall describe in some detail two methods of taking samples from the public high schools of Minnesota. The total number of such schools ($N = 496$) have been classified by (1) type and (2) size of enrollment. We wish to select a sample of 100 schools to be

[1] Palmer O. Johnson, *Statistical Methods in Research*, Chapter IX (New York: Prentice-Hall, Inc., 1949).

used for assembling data by interview and from the school and tax records for the purpose of making a detailed study of the cost of instruction per individual student. Method 1 is that of the unrestricted random sample. Method 2 is that of the proportional or variable fractional sample.

1) *Method of the unrestricted random sample within strata.* Suppose we desire to take an unrestricted random sample of 100 from the 496 high schools of Minnesota. We have these schools classified in a bivariate system, i.e., size of school by type of school. The procedure to be followed in taking this sample will now be explained.

Step 1. Enumerate your population. As the individual members are not given, it is sufficient to cumulate the frequencies in the cells of Table 13.

Step 2. On a separate sheet of paper list these frequencies. For example, the 97 four-year high schools of enrollments 25–74 were assigned the numbers 3–99.

Step 3. Open Fisher and Yates' *Statistical Tables for Biological, Agricultural and Medical Research* to the Random Number Table at a preselected row and column and start enumerating in a direction also preselected.

Step 4. As the numbers in this random number table are given in pairs we proceed by considering two sets of the pairs. Our cumulative frequencies do not exceed 496, so we ignore all random numbers greater than 0496. Consequently, enumeration proceeds by running down (say) two columns and tallying 100 numbers which fall below 0497 for the

TABLE 13. Classification of the 496 Public High Schools of Minnesota by Type and Size of Enrollment

	Size of High School *										
Type	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*f*	*cf*
High School Department	2 0–2	0 —	0 —	0 —	0 —	0 —	0 —	0 —	0 —	2 —	2
Four-Year High School	97 3–99	19 100–118	9 119–127	8 128–135	7 136–142	3 143–145	7 146–152	3 153–155	1 156	154	156
Six-Year High School	84 157–240	55 241–295	31 296–326	27 327–353	6 354–359	0 —	0 —	0 —	0 —	203	359
Junior-Senior High School	0 —	3 360–362	11 363–373	27 374–400	50 401–450	23 451–473	6 474–479	4 480–483	0 —	124	483
Senior High School	0 —	0 —	0 —	0 —	0 —	1 484	3 485–487	4 488–491	5 492–496	13	496
	183 183	77 260	51 311	62 373	63 436	27 463	16 479	11 490	6 496	496	

* Size of enrollments:

1) 25–74	4) 125–174	7) 600–974
2) 75–99	5) 175–349	8) 975-1749
3) 100–124	6) 350–599	9) 1750 and over

list made in Step 2. Thus the 16 four-year high schools with enrollments from 25 to 74, inclusive, (see Table 14, Col. 1, row 2), represent the 16 of the total random numbers examined, as they occurred with values between 003 and 099, inclusive.

Step 5. Total the tallies and enter them in new table (Table 15). The number in each cell represents then the number of individual schools

TABLE 14. A Sample of 100 Schools Drawn by the Method of Unrestricted Random Sample Within Strata

	Size of High School									
Type	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Total*
High School Department	0	0	0	0	0	0	0	0	0	0
Four-Year High School	16	2	1	1	0	0	2	0	0	22
Six-Year High School	13	17	11	7	1	0	0	0	0	49
Junior-Senior High School	0	0	3	5	11	4	2	0	0	25
Senior High School	0	0	0	0	0	0	0	2	2	4
Total	29	19	15	13	12	4	4	2	2	100

from the total number in the population that were chosen by the random process employed.

2) *Method of proportional or variable fractional sample.* It should be noted that the method is valid for taking an unrestricted random sample for this case. No attempt has been made to make the sample frequencies proportional to the population values. Such a procedure would alter the technique above but would ensure a

TABLE 15. A Sample of 100 Schools Drawn by the Method of Proportional or Variable Fractional Sample

	Size of High School									
Type	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Total* *
High School Department	(.40 = f_e) 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	(.40) 0
Four-Year High School	(19.55) 20	(3.83) 4	(1.81) 2	(1.61) 2	(1.41) 1	(.60) 1	(1.41) 1	(.60) 1	(.20) 0	(31.02) 32
Six-Year High School	(16.93) 17	(11.08) 11	(6.25) 6	(5.44) 5	(1.21) 1	0 0	0 0	0 0	0 0	(40.91) 40
Junior-Senior High School	0 0	(.60) 1	(2 22) 2	(5.44) 5	(10.08) 10	(4.64) 5	(1.21) 1	(.81) 1	0 0	(25.00) 25
Senior High School	0 0	0 0	0 0	0 0	0 0	(.20) 0	(.60) 1	(.81) 1	(1.01) 1	(2.62) 3
*Total * Expected* f_e	(36.88)	(15.51)	(10.28)	(12.49)	(12.70)	(5.44)	(3.22)	(2 22)	(1.21)	(99.95)
Rounded	37	16	10	12	12	6	3	3	1	100.00

* Further calculation would show that the sums on the margins of Table 13 multiplied by $\frac{100}{496}$ would reveal slight discrepancies with those marginal values of Table 15. This is due to our "rounding" process.

more representative sample. A suggested method for stratification would be to multiply each population frequency by $\frac{100}{496}$ and enter the results as the number of sample values needed. A rounding off procedure should be followed. The exact schools of the sample can be determined almost identically by Steps 1–5, but we must ignore members which exceed the expected frequencies as described in Step 4. (An observing student would note that this method is somewhat similar to the method of determining the expected values, f_e, in an r x c fold table in calculating χ^2.) The expected frequencies together with the rounded sample values are given in Table 15.

Sampling with unequal probabilities. A method of sampling which uses stratification as a device for allotting unequal probabilities of selection to different population elements[1] can be illustrated by the situation in which we might wish to determine norms for the state of Minnesota on achievement tests to be used in the public high schools. Here we wish to estimate the average score on each of these tests for all students attending the public secondary schools of the state. Funds and time available for the study may require that the sample be restricted to a limited number of the students from schools within the school systems. The following procedures could be used in drawing the sample, whose size has been determined by the maximum accuracy obtainable from the funds available.

1) The public secondary school systems of the state are first grouped into strata. The strata are established so as to contain approximately equal numbers of students. It is desirable to obtain strata as nearly homogeneous as possible with respect to achievement. Some of the strata may contain only one school system. It is planned to draw a single school system from each stratum, so the number of strata should equal the number of school systems which are to be drawn.

2) Within each stratum each school system is assigned a probability of selection. The probability of selecting a school system may be made proportional to the number of secondary students enrolled. Sets of random sampling numbers may be used for assigning unequal probabilities. Thus a block of consecutive numbers equal to the enrollment in the school can be assigned to each school system to be drawn.

3) A similar technique may be used for selecting schools within each of the selected school systems. Thus each school in the system is given

[1] Eli S. Marks, "Some sampling problems in educational research," *The Journal of Educational Psychology*, 1951, 42:85–95.

a probability of selection proportional to the entire enrollment.
4) Finally, students within the selected schools can be selected at random, making use of the school's roster of students. The proportion of students to be sampled from a given school can be established such that every public secondary school student in the state has the same probability of being included in the sample. This procedure makes it unnecessary to use weights in preparing the sample estimates.

7) **Treatment of the nonrespondents.** In many types of survey, there are a number of units in the sample from which the information sought cannot be secured at the first attempt. This may result from inability to make contact with respondents who may not be at home as in the case of the interview or who do not reply as in the case of a mail questionnaire. The "nonresponse" group constitutes an important practical problem. Unless the nonresponse group constitutes a very small proportion of the whole sample, the results obtained are invalid. To obtain the delinquent information may require much time and money; but every effort should be made to reduce the nonresponses to negligible proportions. To ignore this group may result in a sample that has a bias of unknown magnitude. A rigorous plan of dealing with the problem of nonresponse must, therefore, be inaugurated at the outset of the sample inquiry.

In the sample-survey interview the number of deliberate nonrespondents is usually small; the number of individuals who failed to respond at first call, however, may be substantial. There is no adequate way of dealing with this number except by persistent callbacks until complete coverage is obtained. Nonresponse is usually most serious in mail-questionnaires. Delays in response are also at times very troublesome, especially if the returns to be of value must be obtained quickly. The first step is to send a follow-up request by letter. If this does not produce results, then the possibility of introducing more forceful methods must be considered. These methods may be, for example, telephone calls, telegrams, personal visits, or an arrangement with someone in the region to visit the nonrespondents.

If the required information cannot be secured for the nonrespondents of the population under survey, then no amount of statistical ingenuity is capable of providing the investigator with information that is certainly representative of the entire population. However, means can be used to indicate how far the excluded portion of the population is similar to the remainder. It is also sometimes possible to make some allowance for lack of similarity.

A subsample can be taken of those not contained in the first (or subsequent) calls and information sought for the more limited number in the subsample, which, if obtained, is used in weighting the subsample in the final results. When a follow-up is to be carried out on a subsample of the nonrespondents, some simple sampling method such as taking every *k-th* nonresponse may be used. Where there has been a good response to the follow-up, initial nonrespondents subsequently responding can be handled as a subsample of all initial nonrespondents and weighted accordingly in the final analysis.

A unique application to the problem of nonresponse has been made of the idea of stratified sampling by Hansen and Hurwitz.[1] For the general formulas developed, the reader is referred to the original article but the principle used may be briefly described. Questionnaires are mailed out in excess of the number expected to be returned and these are followed up by interviewing a sample of those who do not respond to the postal inquiry. Thus, in the simplest case, the first step is to take a random sample of n units. Of these let n_1 be the number that respond and n_2 the number of nonrespondents. By repeated efforts, the information is later secured from a random sample of r_2 out of the n_2. If

$$n_2 = ar_2$$

the quantity a is the ratio of the sampling rate in the first stratum to the rate in the second stratum. The values of n, the initial size of the sample, and a are selected in order to ensure a designated accuracy for the lowest cost.

The minimum cost of the survey is calculated for each response rate. From this calculation it is possible to specify the maximum number of schedules to be mailed independent of the rate of response. Then to obtain the desired accuracy, the number of individuals to be interviewed would vary with the response rate that was actually found.

8) Conducting the pilot or exploratory survey. The desirability of pilot or exploratory surveys arises from the fact that there are many points in the planning stage of an inquiry by sample on which decisions can be effectively made only after preliminary investigations in the form of a pilot survey have been conducted. With large scale surveys particularly dealing with populations concerning which there is little or no prior knowledge a preliminary exploratory survey may be needed even before an appropriate pilot

[1] Morris H. Hansen and William N. Hurwitz, "The problem of non-response in sample surveys," *Journal of the American Statistical Association*, 1946; 41:517–529.

survey can be attempted. Even with a small-scale survey, a pilot study may yield information of importance in planning the investigation, particularly if it deals with a population about which nothing is initially known.

The main purposes of a pilot survey are (1) to provide information on the various components of variability in the material to be investigated, and (2) to develop and test field techniques, to try out questionnaires, and to provide investigators training and experience. The pilot survey may also provide useful data in estimating costs of the different operations, in determining the most effective type and intensity of sampling and size and number of sampling units.

When the questionnaire is used, it should be tested on a small representative sampling in the field before the survey begins. Such pretesting should indicate questions that are ambiguous or not clearly worded, questions that are difficult for the respondent, and such queries as the respondent may have with respect to the meaning of certain questions. The sample returns may also be used to estimate the optimum number of questionnaires to be distributed for a desired precision, and the expectation of the number of nonrespondents as a basis for laying plans for procedures to obtain as complete coverage as possible. Likewise, when new methods of observation and measurement are to be used they should be tried out on a more or less random sample of the population before being set into operation.

Yates [1] illustrates how modern methods in the design of experiments can be used to make rigorous tests of different forms of the same question with respect to the answers received, or to test differences between investigators using the same form of questions. For example, if two forms, say *P* and *Q* of a question and three field investigators A, B, and C are to be tested, groups or *blocks* of six respondents may be used. Using available prior information, the blocks are selected so as to make the respondents within each block as similar as possible. There are six question-investigator combinations: *P*A, *P*B, *P*C, *Q*A, *Q*B, *Q*C. These should be assigned at random to the respondents of each block. The technical name given to the form of experimental design is a 2 x 3 *factorial design in randomized blocks*. By this design differences between forms of question and between investigators are tested at the same time and information is also obtained concerning the interactions between

[1] Frank Yates, *Sampling Methods for Censuses and Surveys* (London: Charles Griffin and Company, Limited, 1949).

forms of question and investigators; that is, whether differences between investigators are different for the different forms of question, and vice versa.

Investigations of this type can be conducted in the actual interview survey. They are more valuable, however, in a particular survey when they are previously introduced as a part of the pilot survey.

9) Summary and analysis of the data. The statistical analysis of the data of sampling surveys is principally based on counts of numbers of units that fall into different classes and subclasses. Where quantitative variates have been accorded, totals for the classes are secured. The units are usually the sampling units but may be at times some other natural unit.

From these numbers and totals the arithmetical means can be computed for the different classes. Basic summary tables can then be compiled. In these tables, the frequencies resulting from the counts are expressed in percentages, the basis of which are chosen so as to show the differences in percentages that are of interest. More critical analysis can now be applied to the data in the summary tables. Usually this further analysis is needed so that the effects of the various factors on which data were collected and which are believed to influence the results may be isolated.

In sample surveys, estimates of sampling errors are required. These provide the bases for needed tests of significance and for problems in estimation which include the setting up of confidence intervals. In addition, investigation of the relative efficiency of different sampling methods may be conducted.

Often there is in addition to the numerical data, important qualitative information which may not lend itself to statistical summary.

Depending upon the extensiveness of the survey data and upon the nature of the material collected, the handling of the data usually takes one of the following forms: (1) analysis direct from the survey forms, (2) transference of the data to ordinary cards, (3) the use of specially prepared cards with holes around the edges, and (4) the use of punched cards. The details of these methods cannot be given here but are discussed by Yates. The principles of punch-card machine operation are also given by Hartkemeier.[1]

The nature of the inferences possible from survey data has been described earlier. They were contrasted with inferences from experimental data. The special value of the survey was pointed out

[1] H. P. Hartkemeier, *Principles of Punch Card Machine Operation* (New York: Thomas Y. Crowell, 1942).

in situations in which experiment was difficult or impossible and as preliminary to experimental work. To be effectively used for either of these purposes the influences of as many extraneous factors as possible must be eliminated. This is essentially the problem of survey design and statistical analysis.

10) Preparation of the sampling survey report. We may conclude this discussion of procedures for the sampling survey by specifying certain points that the investigator should include in his report, insofar as they apply to his particular case.[1]

(1) *General description of the survey.*
  (a) Statement of purposes of the survey.
  (b) Description of the material covered.
  (c) Nature of the information collected.
  (d) Method of collecting the data.
  (e) Sampling method.
  (f) Accuracy.
  (g) Repetition.
  (h) Point or period (of time).
  (i) Date and duration.
  (j) Cost.
  (k) Responsibility.
  (l) References.
(2) *Design of the survey.*
(3) *Method of selecting sample-units.*
(4) *Personnel and equipment.*
(5) *Costs.*
(6) *Accuracy of the survey.*
  (a) Precision as indicated by the random sampling errors deducible from the survey.
  (b) Degree of agreement observed between independent investigators.
  (c) Other nonsampling errors.
  (d) Accuracy, completeness, and adequacy of the frame.
  (e) Comparison with other sources of information.
  (f) Efficiency.

ILLUSTRATIONS OF SAMPLING SURVEYS

Finally, we shall report briefly on two investigations in education that have used the sample-survey-method.

[1] These points and others are reported more fully by Yates and are based on the memorandum prepared by the United Nations Sub-Commission of Statistical Sampling, entitled, *Recommendations Concerning the Preparation of Reports on Sampling Surveys*. See also "The preparation of sampling survey reports, Statistical Office of the United Nations," *The American Statistician*, 1950, 4:6–12.

Cornell's[1] study was prompted by a need of information basic to certain administrative decisions concerning several nation-wide college and university programs. The basic information about the expansion of enrollments was required as soon as possible after the fall opening. Unbiased estimates were needed of total enrollments of various types of students in each of six major classes of higher educational institutions. A degree of precision was specified whereby the estimates were to have coefficients of variation of .05 for each of the six classes. The requirements of the sample survey, therefore, included speed, accuracy, and advance knowledge of reliability. The method of stratification was used; type of institution and size data from previous enrollment records provided the basis of classification. The principle used for allocating sampling units to the respective strata is that the number of sampling units in a stratum is proportional to the product of the size of the stratum and the standard deviation of the stratum. This procedure makes the representation of the stratum in the sample compatible with both its size and variability.

The estimates, sampling errors for estimates, and coefficients of variation reported by Cornell are given in Table 16.

Anderson contributed a unique study in educational evaluation of the extent to which certain specified objectives in the fields of biology and chemistry were achieved during an academic year in the public secondary schools in the state of Minnesota.[2] He first selected a representative sample of 56 high schools in the state making use of the principle of proportionate sampling from schools stratified according to population centers. The total number of pupils involved in the biology portion of the study was 1,980; in chemistry, 1,352. An intelligence test and pretests for each of four objectives were administered to all the pupils in September and final tests for the same objectives at the end of the school year in May. Each co-operating teacher kept throughout the school year a detailed log which provided descriptions of procedures and materials of instruction. In addition, various teacher characteristics and pupil characteristics were studied. Through extensive use of the techniques of analysis of variance and covariance he made fourteen comparisons in biology and fifteen in chemistry which resulted in the identification of those teacher characteristics, pupil factors, and teaching procedures which were significantly associ-

1 Francis G. Cornell, "Sample plan for a survey of higher education enrollment," *Journal of Experimental Education*, 1947, 15:213–218.

2 Kenneth E. Anderson, "A frontal attack on the basic problem in evaluation: the achievement of instruction in specific areas," *Journal of Experimental Education*, 1950, 18:163–174.

TABLE 16. Sampling Errors for Estimates of 1946 Fall Enrollment in Higher Educational Institutions

Type of Institution	*All Students*			*Men*			*Women*			*Veterans*		
	S	$\hat{\sigma}_s$ Amount	C.V. Per Cent	S	$\hat{\sigma}_s$ Amount	C.V. Per Cent	S	$\hat{\sigma}_s$ Amount	C.V. Per Cent	S	$\hat{\sigma}_s$ Amount	C.V. Per Cent
All institutions	2,078,095 *	46,750 †	2.2 ‡	1,417,595	40,152	2.8	660,500	16,679	2.5	1,080,396	33,159	3.1
1. Universities and large institutions of complex organization	1,031,430	37,515	3.6	762,423	32,063	4.2	269,007	11,714	4.4	591,468	25,825	4.4
2. Colleges of arts and sciences	439,449	22,883	5.2	249,738	20,851	8.3	189,711	9,935	5.2	194,570	18,434	9.5
3. Independent technical and professional schools	210,176	7,067	3.4	178,409	6,134	3.4	31,767	2,917	9.2	129,238	5,396	4.2
4. Teachers colleges and normal schools	150,059	6,290	4.2	78,963	5,180	6.6	71,096	3,306	4.7	61,780	4,289	6.9
5. Junior colleges	188,139	12,276	6.5	121,069	8,848	7.3	67,070	4,486	6.7	85,124	6,456	7.6
6. Negro institutions	58,842	3,786	6.4	26,993	2,573	9.5	31,849	1,645	5.2	18,216	1,900	10.4

* Estimated enrollment.
† Unbiased estimate of population standard deviation, or standard error.
‡ Coefficient of variation = σ_s/S.

ated with student achievement. The survey evidence provided a valid basis for generalizations characterizing the effectiveness of teaching in biology and chemistry in the state.

An illustration of the type of analysis carried out is given by the analysis of variance and covariance[1] tables dealing with the evaluation of the achievement of students in biology in ac-

TABLE 17. Analysis of Variance and Covariance of Final Scores with Otis Score and Pre-Test Score Constant—Scores of Students Taught by Teachers With Different Amounts of Training in Science

Source of Variation	*D.F.*	*Sum of Squares*	*Mean Square*	*F*	*Probability*	*Hypothesis Tested* *
Within groups	174	12082.0802	69.4372			
Between groups	1	591.7627	591.7627	8.52	P < 01	Rejected
Total	175	12673.8429				

* The null hypothesis, that there was no difference between the means holding intelligence test score, and pre-test scores constant was tested.

cordance with the amount of training in science which the teachers had had. In this comparison in biology, there were thirteen teachers who had taken 77 or more quarter hours of science in college (group A) and thirteen who had taken 32 or less quarter credit hours of science in college (group B). See Tables 17 and 18. From Table 17 it is noted that a significant difference between the adjusted mean achievement of group A and B exists. As shown in Table 18, the group A had an adjusted mean of 50.19 as compared with one of 46.06 for group B. That is, students in biology classes taught by teachers with 77 or more quarter hours of college science, achieved on the average significantly more than did pupils taught by teachers with 32 or less quarter hours of college science. Through the application of the technique of analysis of covariance, the final means were adjusted for whatever inequalities existed in the two groups with respect to intelligence tests and pretest scores.

TABLE 18. Adjusted Method Means—Comparison

Group	*Mean*			*Difference from Grand Mean*		b_1	b_2	*Corr. b_1*	*Corr. b_2*	*Adjusted Mean*
	Otis	*Pre-Test*	*Final*	*Otis*	*Pre-Test*					
Group A	43.703	32.890	49.726	−.142	1.014	.592652	.536072	−.084	.54	50.19
Group B	43.200	36.500	47.240	.361	−2.596			.213	−1.392	46.06
Grand Mean	43.561	34.904								

SUMMARY

Within recent years sampling has been increasingly used to ascertain information necessary in answering certain questions about a specific population or populations. At times the problem that prompted the investigation may be called an enumerative problem; at other times an analytical one. In the former, the object of the sample survey is to determine or estimate certain characteristics of the population without inquiring into the reason as to why these characters are what they are. In the analytic problem the purpose of the sampling survey is to inquire into the underlying factors or causes that may have given rise to an observed condition or situation. The sampling survey provides information with respect to the prevalence of conditions that are presumed to give rise to certain effects. The actual establishment of these relationships must be done by experimental research. The ultimate purpose in studying causes is to be able to predict the results of certain causes with a view to their control. It may be noted that the collection of enumerative data may provide important source data for analytical investigations.

Although experiments are essential for determining the magnitude of the effect of any factor or combination of factors, surveys are of particular value where experiments are difficult or practically impossible. They are also important in furnishing preliminary information by identifying factors that are promising for experimental investigation.

If sample surveys are to provide the basis for decision and action, the sample results must be capable of translation and interpretation in such a way that may provide maximum information. The conclusions are drawn for the population. They are inferred from the information available in the sample results. To provide this basis for decision and action the sample must be a probability sample.

The theory of probability cannot be applied to a sample that is not randomly taken. Hence it is not possible to measure the degree of confidence to be placed in any inference drawn from a nonrandom sample. In addition to the random process of selection, the biases of selection, nonresponse, and estimation must be in effect eliminated or contained within known limits. This is necessary if sample errors are to be calculated and if probability statements involved in testing statistical hypotheses and in estimating population values are to have meaning.

CHAPTER VII

Search for Interrelationships

In solving many of the complex problems with which field workers are concerned, there will be a continuous shifting from status to relationships and from relationships to status. In studying status one is obviously led to consider factors producing that status and to the interrelationships of such factors. For example, a typical survey of school achievement leads to consideration of the factors underlying achievement, their importance, and interrelationships. This interplay between status and interrelationship is emphasized by questions such as the following: What conditions underlie status? How have these conditions come to be? How do they appear in different situations?

The many studies of relationships may be classified into two large categories according to the symbolism employed, namely the nonmathematical and the mathematical. The first of these depends upon logical first principles without much statistics, such as case studies, comparative studies, and historical studies. The second is primarily statistical in character, such as correlational studies, analysis of variance, and related techniques. Our concern in this chapter is with the nonmathematical studies of relationships.

THE CASE STUDY

The case study is potentially the most valuable method known for obtaining a true and comprehensive picture of individuality. It makes possible a synthesis of many different types of data and may include the effects of many elusive personal factors in draw-

ing educational inferences. It seeks to reveal processes and the interrelationships among factors that condition these processes.

Initial concepts in new fields of science frequently result from the analysis of individual cases. Before the field of psychiatry asserted itself as a science, the behavior of a disturbed individual had to be regarded as constituting a novel problem. Almost no categories existed under which to place symptomatic manifestations of an individual's behavior, and almost nothing was known about the probability with which specific symptoms might be intercorrelated. Study of large amounts of descriptive data derived from observation of many individuals eventuated through application of principles of agreements and differences, in a formulation of what may for convenience be called "types"of disorder. Assuming that a given "type" could represent a definable array of symptoms, psychiatrists began to speak a common language; and communicable research in the field became possible.

After certain concepts have become available for giving direction to research, investigation can take the form of studying intensively a relatively small array of personal characteristics but doing so with vastly increased samplings of population. In addition to learning much about *one* individual's composite of characteristics, as in the case of Jones' genetic study [1] of an individual, the research worker also seeks verification of hypotheses by gathering small but significant amounts of information about *many* individuals. In the succeeding sections of this chapter much will be said about comparative studies. An excellent source of data for comparative studies will be found in case studies.

Clinicians, educational and vocational counselors, and school psychologists are continually gathering case study data in their dealings with individuals. Although their interests are with an individual *per se*, information developed in such fields, is an extremely valuable source of data for many types of research. The fact that the clinician is professionally interested in the treatment of individuals, and to a less extent interested in developing generalizations, should not detract from the use of their data.

A significant instance of such a study is that of gifted children conducted during two decades by Terman [2] and his associates. A

[1] H. E. Jones, *Development in Adolescence* (New York: Appleton-Century, 1943).

[2] L. M. Terman, *Genetic Studies of Genius:* Vol. I., *Mental and Physical Traits of a Thousand Gifted Children* (Stanford University Press, 1925); C. M. Cox, Vol. II., *The Early Mental Traits of Three Hundred Geniuses* (Stanford University Press, 1926); B. S. Burks and others, Vol III., *The Promise of Youth* (Stanford University Press, 1930); L. M. Terman and M. H. Oden, Vol. IV., *The Gifted Child Grows Up* (Stanford University Press, 1947).

considerable amount of case study history related to children of exceptionally high intelligence was used in this study. The study was undertaken to determine the characteristics of such individuals as a class rather than to deal with them as clinical problems. It was continued during the lives of many of the individuals concerned so long as it was possible to maintain contacts with them.

Harvey[1] compared the backgrounds of socially maladjusted Mexican and American boys through an analysis of a large number of individual case studies. The comparison was made with reference to three major areas, namely, socio-economic, physical, and psychological factors. A summary of the socio-economic data is presented in Table 19. The author comments upon these as follows:

> The delinquent Mexican boy, born in Los Angeles, is the product of a foreign family background whose religion is predominantly Catholic. The majority of these boys have a bilingual handicap. The parents of the Mexican boys had little education; the majority of the American parents attended or graduated from high school. A classification of the marital status of the delinquent boys' parents revealed that over half of the parents, both Mexican and American, were divorced or separated, although death was the main cause for the broken homes in the Mexican families and divorce led as a causal factor in the American homes. The average size of the Mexican family was greater than the American family. The mean number of siblings in seventy-one Mexican families was found to be 5.0 as compared to 3.2 for seventy-five American families. The average size of the Mexican family was 6.8 and that of the American group was 4.9. Half of the Mexican families fell into the marginal economic classification, as may be noted by the percentage figures in Table 19, whereas the average American family was classified in the moderate or good economic categories. When the monthly incomes from all sources for the Mexican and American families were tabulated, it was found that the average monthly income for the Mexican family was $185.71 as compared to $253.70 for the American family. Although the difference in monthly income between the two groups was $77.99, it is significant that two more people, on the average, were living on the Mexican family's income than were living on that of the American family. This factor increased the difference in economic status between the two groups more than monthly income alone. In respect to economic classification [the table], reveals that, whereas more than half of the parents of the American boys were noted as professional, semi-professional, or skilled workers, the Mexican parents were usually unskilled laborers.

[1] Louise F. Harvey, "The delinquent Mexican boy," *Journal of Educational Research*, 1949, 42:573–585.

TABLE 19. A Comparison of Some of the Socio-economic Data Obtained Relative to the Families of the Delinquent Mexican and American Boys

Socio-Economic Data	Mexican Number	Families Per Cent	American Number	Families Per Cent
1. MARITAL STATUS	75	100.0	75	100.0
Unbroken homes	33	44.0	24	32.0
Broken homes	42	56.0	51	68.0
2. ECONOMIC STATUS	60	100.0	72	100.0
On relief	11	18.3	4	5.6
Marginal	30	50.0	12	16.6
Fair	12	20.0	8	11.2
Moderate	5	8.3	34	47.2
Good	2	3.4	14	19.4
3. JOB CLASSIFICATION	52	100.0	64	100.0
Professional	0	0.0	2	3.1
Semi-professional	0	0.0	6	9.4
Skilled workers	8	15.4	32	50.0
Laborer	41	78.8	22	34.4
Small business	3	5.8	2	3.1
4. CLASSIFICATION OF HOME ADDRESSES BY NUMBER OF BLIGHTING INFLUENCES	75	100.0	75	100.0
Area I (1–4)	0	0.0	19	25.3
Area II (5–8)	1	1.4	13	17.3
Area III (9–12)	16	21.3	2	2.7
Area IV (13–16)	13	17.3	0	0.0
Area V (above 16)	22	29.4	1	1.3
Not in city	20	26.6	39	52.1
Unknown	3	4.0	1	1.3

A classification of the home addresses of the families of the delinquent boys was made to determine the types of areas in which they lived. [The table] gives the result of the investigation. It was found that the Mexican families, on the whole, lived in areas of relatively high blight as compared to the families of the American delinquents. An interesting aspect of the sampling, in the American group, was that half of the American delinquents came from the suburbs around Los Angeles or from other sections; but 68.0 per cent of the Mexican boys came from the blighted areas of the city of Los Angeles. Mexican boys who lived in Belvedere, a blighted area on the outskirts of the city, were not

included and they amounted to 17.0 per cent more of the group. Degree of blight was determined by the number of blighting influences which were rated in relation to: low average rents, high juvenile delinquency, low assessed value of land, dwellings needing major repairs, and lack of sanitary facilities. Area I, for example, had a rating of from one to four, and each area had an increasing number up to Area V, which had above sixteen blighting influences. The data were found by locating the home address of each delinquent boy on a map of Los Angeles and relocating the section on a map of blighted areas prepared by the Los Angeles City Planning Commission in 1945.

Ackerson's[1] study of children's behavior problems is another study based upon analysis of case study records. It involved the study of 2,113 boys and 1,181 girls between ages six to eighteen selected from 5,000 consecutive cases examined at the Illinois Institute for Juvenile Research. The general procedure in quantifying significant aspects of case records was that of classifying the types of problem, tabulating frequency of occurrence under each category (differentiating by sex), and presenting as results the most significant of correlations calculated between various types of difficulty in which the children were involved.

A portion of a typical table[2] (see Table 20) is presented in order to illustrate the general nature of the information as quantified and refined by statistical treatment.

TABLE 20. Correlations with "Stealing"

	Boys	*Girls*
Personality—total	.19 ± .02	.29 ± .03
Conduct—total	.55 ± .01	.51 ± .02
Police arrest	.63 ± .02	.37 ± .02
Truancy from home	.64 ± .02	.54 ± .03
Truancy from school	.62 ± .02	.39 ± .04
Lying	.61 ± .02	.68 ± .02

In the complete table, sixty-three items were correlated with "stealing." Only some of the larger positive correlations are here presented.

The validity of conclusions in studies based upon case study may be influenced by (1) effect of selection upon the representative quality of the sampling, (2) individual bias on the part of informants or examiner, (3) variability in completeness in certain areas of the data, and (4) inadequate definition of categories. When case

[1] L. Ackerson, *Children's Behavior Problems*, Vol. II, *Relative Importance and Intercorrelations among Traits* (U. of Chicago Press, 1942).

[2] L. Ackerson, *op. cit.*, pp. 312–313.

studies involve a "problem" population, the individual's undesirable traits are usually overemphasized and his desirable traits underemphasized. Such a study may be of questionable validity in "normal" populations.

Steps in a case study. Generally in conducting a case study one goes through the following steps:

1) *Establishes* the fact that the phenomenon under investigation, frequently an individual, is inadequate in some vital respect.
   a) Collects what appears to be relevant data, observes behavior, administers tests, examines products.
   b) Evaluates the data collected, compares data with past experience and norms.
   c) Reaches a decision that not all is well; that the conditions leading to or accompanying the inadequacy must be sought and remediation applied.

2) *Selects* from among the circumstances leading to or accompanying the observed inadequacy a supposed cause or causes.
   a) Reviews his own past experience, consults with others, and re-examines the scientific literature relative to similar situations.
   b) Looks for symptoms that might indicate the presence of some disabling deficiency.
   c) Formulates hypotheses about the probable causes of the deficiency observed.
   d) Checks for the presence or absence of the supposed cause, through systematic investigation when such appear necessary.

3) *Institutes* a remedial, corrective, or improvement program.
   a) Re-examines his own past experience and scientific investigations for ideas relating to a course of action.
   b) Chooses from several alternate courses of action those that appear to be appropriate to the immediate situation.
   c) Institutes a treatment program.

4) *Rechecks* to determine adequacy of behavior, performance, or output.

The movement is not as straightforward as the steps listed above would suggest, but there is an inductive-deductive process wherein a number of aspects of the situation are attacked sometimes more or less simultaneously, with increasing intensity. As the situation clarifies, it tends, however, to assume logical steps such as those suggested above.

We have referred to case studies in the preceding discussion in very general terms. The study may concern itself principally or solely with currently discoverable symptoms and causes as in a medical diagnosis, or with case histories. To comprehend complex cases one will ordinarily carry the study beyond the present to evidence of past inadequacies. Though the basic logic and purposes may be much the same in either case, the distinction between case history and current diagnosis may be appropriately maintained.

Before continuing with a more detailed discussion of the diagnostic process, it may be interesting to note the many points at which status studies are conducted in making case studies. One begins a diagnostic study by noting that the status of some individual is inadequate in some vital respect. In the search for potentially disabling disabilities one conducts many qualitative and quantitative status studies, as one does in checking upon the presence or absence of hypothecated disabilities. If a remedial program is instituted, one ascertains its effects by further studies of status. Our interest at this point is, however, in the discovery of interrelationships.

Diagnostic studies may be conducted in many fields of specialization. The term diagnosis, as has already been stated is widely used in education, psychology, and medicine. Diagnostic techniques are applied in such fields of specialization as psychiatry, psychometrics, sociometrics, psychoanalysis, abnormal psychology, psychosomatic medicine, psychotherapy, counseling, personality research, mental hygiene, mental diseases, psychopathology behavior problems, and scholastic difficulties.

Many different types of data-gathering devices are used in diagnostic studies. The data-gathering devices employed in making diagnoses are numerous, including psychological tests, questionnaires, directive and nondirective interviews, projective techniques, sociometric methods, behavior-rating devices, check-lists, and inventories. Each device is represented by a long list of specific instruments. Psychological tests, for example, may be of general ability, special abilities, personality, interests, attitudes, and adjustment. Each of these types of data-gathering devices includes many subtypes. General ability tests, for example, include tests of intelligence, verbal and nonverbal, which may be further subdivided into tests of academic, vocational, social, and aesthetic intelligence. Special-ability tests include tests of reaction time, coordination, comprehension, artistic ability, mechanical ability, aptitudes, and sensory abilities.

The validation of diagnoses. Far too little attention has been given to the validation of diagnoses. The fact that the examiner has great faith in his diagnosis, or that he has gone through a routine established by the mores of his profession, by no means assures diagnostic validity. There must be tangible evidence of validity. Clinical psychologists stress the use of clinical judgment in the interpretation of statistical, historical and interview data. Judgments, even those of experts, that are carefully made, are nonetheless subject to error and must be so regarded. Those using clinical data should constantly seek validating data.

The following suggestions will be helpful in validating clinical diagnoses:

1) All data-gathering devices should be carefully validated: observational devices, interviews, tests, questionnaires, inventories, projective techniques, and mechanical instruments.

2) Every properly validated data-gathering device has an ascertainable sampling error, as well as systematic errors of measurement. The probable (or standard error) plus infomation about errors of measurement should be attached to all scores (as of tests) and judgments (as of clinicians).

3) Generalizations developed from group data may not apply to individuals. The generalization may be valid for the group as a whole but there may be many individual exceptions. The case under investigation may be such an exception.

4) Diagnoses should be based when practicable upon multiple signs data from more than one device, and from more than one examiner secured under different conditions in order to reduce error.

5) Diagnoses that adhere to a systematic plan (for example, as that given above) for collecting and interpreting data are more likely to yield valid results than those based upon general impressions, estimates, and guesses.

6) The ultimate validation of the diagnosis and remedial program will be found in the subsequent life adjustment and performance of the subjects.

Other considerations that should improve the quality of diagnostic studies. In addition to the generalizations relative to the validation of clinical diagnoses, offered above, there are a number of considerations that should improve the quality of diagnostic studies.

First, normal behavior for different individuals under different conditions in different areas of human activity must be defined. If

everyone is not to be under suspicion, and judged by the private systems of values and idiosyncracies of self-appointed evaluators, there must be standards of competency, behavior, and performance. This applies to both qualitative and quantitative judgments about individuals.

Second, the factors believed to limit or facilitate human competency and performance in various areas of human activity should be carefully defined and categorized. Here again many examiners have their own private systems of categorization and definitions. The data collected relative to the same individual may vary widely under individual systems of data collecting.

Finally, a system of carefully validated generalizations should be made available to diagnosticians in all fields.

Halliday,[1] for example, developed the following generalizations as an aid to the study of psychosomatic affection:

1) Emotional tension precipitated by some upsetting event in the patient's life is important in a large proportion of cases.
2) The personality of the patient tends to be associated with certain kinds of disease.
3) A definite disproportion of sex ratio exists in many of the disorders.
4) Other psychosomatic disorders often occur simultaneously or may alternately occur in the patient.
5) A similar or associated disorder in parents, relatives, or siblings is noted in a significantly high proportion of cases.
6) In psychosomatic disorder the course of the illness tends to be phasic with periods of improvement followed by periods of recurrence.

Good diagnoses proceed from systematizations such as these. There must be thorough understanding of the categories to which individuals may be assigned.

Carefully defined categories based upon a well-integrated system of concepts are essential to good case studies. Good diagnoses will not arise from the routine application of data-gathering devices and techniques. They grow out of the richness of experience of the diagnostician plus penetrating insights into the nature of human behavior.

Realizing the importance of well-integrated systems of concepts as aids to practitioners, many students have attempted to systematize the thinking in some field of research. This has been done not only in education but in psychology and medicine as well —all fields in which extensive use of clinical diagnoses have been

[1] James L. Halliday, "The significance of the concept of a psychosomatic affection," *Psychosomatic Medicine,* 1945, 7:240–245.

made. Although excellent summaries have been prepared for a number of fields, the systematizations attempted in personality organization, reading, and the supervision of instruction may be cited to illustrate how complex subjects have been analyzed to provide a suitable frame of reference for case work.

Studies of personality organization. Many persons have attempted systematic categorizations of personality. Statements by Allport, Murphy, and Cattell, among others, may be regarded as illustrative of efforts to systematize the thinking about personality. Cattell [1] after full chapter discussions of (1) the nature and varieties of clinically distinguishable personality forms, (2) description of the principal pathological syndromes, (3) basic methods for delimiting and measuring common and unique traits, (4) systematization of description and measurement scales, (5) the principal surface traits discovered through behavior ratings, (6) the principal source traits discovered through behavior ratings, (7) the principal source traits based on self-inventories, and (8) the principal source traits discovered through objective test measurements, lists the following primary traits about which he organizes the immense literature in this field:

I. Cyclothymia—schizothymia trait
II. Intelligence, general mental capacity—mental defect.
III. Emotionally mature, stable character—demoralized general emotionality.
IV. Dominance—ascendance—submissiveness
V. Surgency—agitated melancholic desurgency.
VI. Sensitive, anxious emotionality—rigid, tough poise.
VII. Trained, socialized, cultured mind—boorishness.
VIII. Positive character integration—immature, dependent character.
IX. Charitable, adventurous cyclothymia—obstructive, withdrawn schizothymia.
X. Neurasthenia—vigorous, obsessional, determined character.
XI. Hypersensitive, infantilesthenic emotionality—phlegmatic frustration tolerance.
XII. Surgent cyclothymia—paranoia.

A systematization such as this is invaluable in providing a suitable frame of reference for personality diagnosis.

Diagnostic studies of reading. In few fields has the work been as extensive and the thinking as generally systematized as in the field of reading. Some idea of the extraordinary amount of re-

[1] Raymond B. Cattell, *Description and Measurement of Personality* (Yonkers, N. Y.: World Book Company, 1946).

search in this field can be observed by an examination of the summaries of research on reading prepared by William S. Gray and reported annually in the February issues of the *Journal of Educational Research.*

Diagnostic studies. Gates [1] outlines a complete diagnostic testing program in reading, an abridged form of which follows:

I. *Reading attainments.*
   A. Word recognition
   B. Sentence reading
   C. Silent paragraph reading
      1. Speed
      2. Accuracy
      3. Power
   D. Oral reading

II. *Reading in context*
   A. Context clues, word-form clues, phonetic devices, etc., in oral reading

III. *Recognition and pronunciation of isolated words.*

IV. *Perceptual orientation and directional habits in reading context.*
   A. Reversal tendencies, omissions of words, failures to observe various parts of words, etc.

V. *Visual perception.*
   A. Ability to work out phonogram combinations
   B. Recognition of various types of word elements, such as
      1. Initial—vowel syllables
      2. Initial—consonant syllables
      3. Vowel—consonant phonograms
      4. Consonant—vowel phonograms
   C. Ability to blend given letters and phonograms into words
   D. Ability to sound individual vowels
   E. Ability to name individual letters

VI. *Auditory perception.*
   A. Ability to spell spoken words
   B. Ability to write words as spelled
   C. Ability to blend letter sounds into words
   D. Ability to name letters when sound is given
   E. Ability to give words with a prescribed initial sound
   F. Ability to give words with a prescribed final sound

VII. *Constitutional and psychological factors.*
   A. Visual perception
   B. Vision
   C. Auditory acuity and discrimination
   D. General intelligence

[1] Arthur I. Gates, *The Improvement of Reading* (New York: The Macmillan Company, 1935). Pages 18–20 reprinted by permission.

E. Memory span
F. Associate learning
G. Muscular co-ordination, handedness, eye dominance, etc.
H. Emotional stability

VIII. *Educational background and environmental influences.*
A. Home conditions
B. School conditions
C. Personal relationships

IX. *Motivation.*

TABLE 21. Major Reading Abilities

Ability	*Rank Order*
I. OBSERVATION	
A. Visual ability	6
B. Auditory ability	18
C. Comprehension	2
D. Speed	4
E. Attention	1
F. Reproduction	3
G. Perception	14
II. RESEARCH ABILITIES	
A. Ability to locate data	12
B. Ability to select data	7
C. Ability to organize data	11
D. Ability to be stimulated creatively	13
III. VOCABULARY ABILITIES	
A. Ability to acquire new vocabulary	5
B. Phonic abilities	15
C. Ability to "unlock" words	8
IV. AESTHETIC ABILITY	
A. Emotional appreciation	9
B. Literary appreciation	17
V. HYGIENIC ABILITIES	16
VI. ORAL READING ABILITIES	10

Burkart [1] in a survey of the literature on reading discovered a total of 214 abilities which reading specialists believed to be involved in the reading process. These many abilities are summarized in Table 21.

For additional illustrative materials the reader is referred to *Clinical Studies in Reading* by the Staff of the Reading Clinics of the University of Chicago.[2]

[1] Kathryn Harriet Burkart, "An Analysis of reading abilities," *Journal of Educational Research,* 1945, 38:430–439.

[2] *Clinical Studies in Reading,* by the Staff of the Reading Clinics of the University of Chicago, Supplementary Educational Monographs (Chicago: University of Chicago Press, 1949), 173 pp.

Barr, Burton, and Brueckner have synthesized materials relating to supervision.[1] Supervision is a very complex activity with many researches treating its diverse phases and activities. These writers have attempted to systematize the extensive literature of this field in the following categories:

1) Determining the objectives of education.
2) Appraising the educational product.
3) Studying the factors limiting and facilitating pupil growth and achievement.
   a) Capacities, interests, and work habits of the pupil.
   b) Teacher factors in pupil growth.
   c) The curriculum.
   d) Materials of instruction.
   e) Socio-physical environment for learning.
4) Improving the conditions that limit and facilitate pupil growth and achievement.
   a) Interests, attitudes and work habits of pupils.
   b) Facilitating teacher growth.
   c) Improving the curriculum.
   d) Improving the materials of instruction and their use.
   e) Improving the socio-physical environment for pupil growth.
5) Evaluating the means, methods, and outcomes of supervision.

Use of profiles. Profiles serve a useful purpose in the summarization of data in case studies. One of the problems in all such studies is that of consolidating the data from various sources in such a way that they can be readily comprehended. The arrangement of data in close proximity to each other as in a profile is a great aid to understanding. It is customary to include on profile sheets not only the data relative to the person, situation, or institution under investigation, but also norms, mathematical or descriptive, that can be used in reaching valid conclusions about the case at hand.[2] The profile in Figure 3 is based upon case studies from data collected through personal interviews.[3] Ratings are provided on this chart for the personality adjustment of two teachers. Each teacher

[1] A. S. Barr, William H. Burton, and Leo J. Brueckner, *Supervision: Democratic Leadership in the Improvement of Learning*, Second Edition (New York: D. Appleton-Century Co., 1947), 879 pp.

[2] See E. F. Lindquist and others, *Educational Measurement* (American Council on Education, Washington, D. C., 1951) for an excellent discussion of precautions to be employed in the use of profiles.

[3] M. Elizabeth Barker, *Personality Adjustments of Teachers Related to Efficiency in Teaching* (Bureau of Publications, Teachers College, Columbia University, 1946), p. 38.

is rated on seven aspects of life adjustment and seven of work adjustment. Case number twenty-three is well above the median on all fourteen aspects of personality for which data are provided. Case number twenty-eight is above the median in only one aspect. Space does not permit the definitions provided by the author for each of the fourteen aspects of personality studied.

Case studies are made for many purposes, two of which have been emphasized: (1) to assess the status of the many factors and their interrelationships that pertain to the well-being of the individual *per se* as employed in guidance and diagnostic work; and (2) to collect individually and personally orientated data useful in the development of generalizations about groups of individuals alike in some vital respect.

Throughout the discussion, our concern has been with the case study as an instrument of scientific investigation.[1] As an instrument of scientific investigation it must satisfy the criteria of science. There must be incontrovertible evidence and not mere supposition, preconceived ideas, and unverified guesses. Where hypotheses are employed they must be tangible, verifiable, and in agreement with the facts. The generalizations must be meticulously drawn. When so employed the case study method can provide data that are invaluable in almost any individually oriented evaluation program.

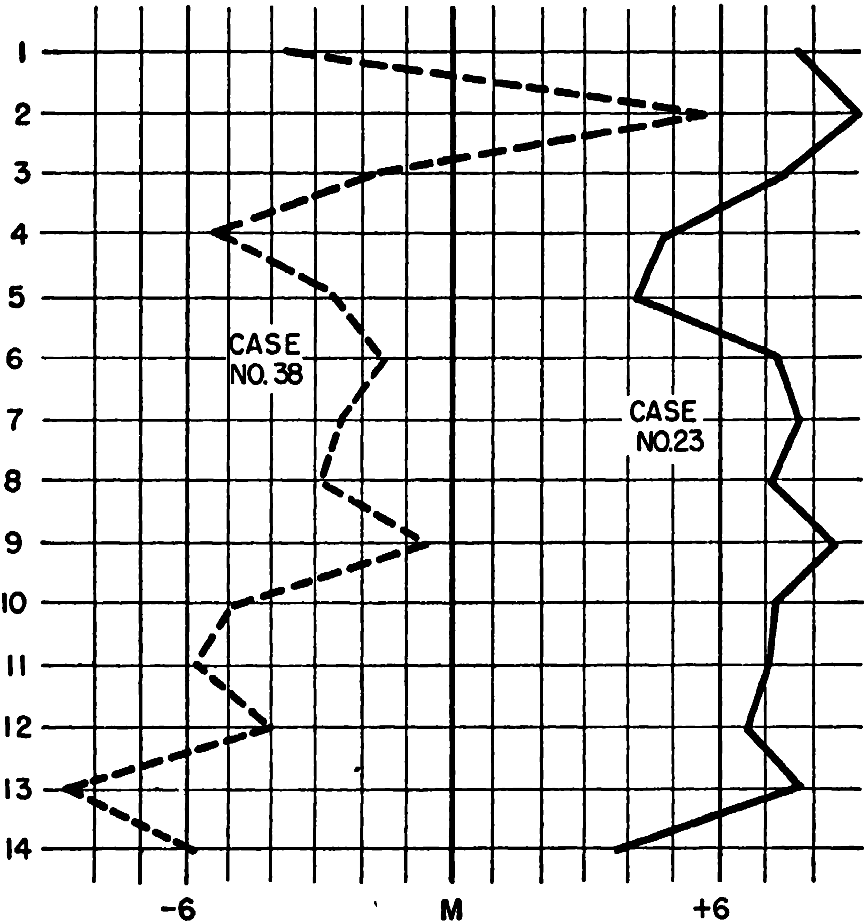
COMPARATIVE STUDIES OF EDUCATION

In a sense, all research and evaluation are comparative. In the analysis of variance, to be discussed later, one employs a statistical procedure for the systematic comparison of variances. In the description and appraisal of status, particularly in the appraisal of status, one compares the status ascertained with like phenomena, criteria, or norms; in correlational studies one compares two or more sets of scores for some particular group of individuals.

We are here concerned however with a special kind of comparative study wherein comparisons are made of products, processes, or conditions without much investigational machinery. The comparisons may be qualitative or quantitative, but the techniques employed are usually nonexperimental and nonmathematical except for statistics of the simpler type.

Two types of studies will be discussed: (1) comparisons of educational practices, processes, and products of various cities, states,

[1] Theodore R. Sarbin, "Clinical psychology—art or science," *Psychometrika*, 1941, 6:391–400.



PERSONALITY ADJUSTMENT KEY

"Life Adjustments"	*"Work Adjustments"*
1. Family relationships	8. Philosophy
2. Living conditions	9. Future goals
3. Health	10. Pupils
4. Finances	11. School administrators
5. Friends	12. School environment
6. Sex	13. Emotional situations
7. Religion	14. Professional growth

FIGURE 3. Personality adjustment of a superior and a below-average teacher.

regions, and counties to ascertain likenesses and differences; wherein the ultimate justification for various phenomena is sought in their historical background, social foundations, and philosophical implications; and (2) comparisons of the *circumstances* accompanying phenomena: products, processes, or practices to discover likenesses and differences among them, the ultimate test of the potency of a given circumstance being sought in appeals to the principles of logic. The first type of study is usually referred to as a comparative education study; the second type of study has usually been referred to as a comparative-causal study.

The comparative-causal type of investigation. The comparative-causal type of investigation finds its justification in the logical principles of agreement and double agreement. These principles may be stated as follows:

1) *The principle of agreement.* If two or more instances of the phenomenon under investigation have only one circumstance in common, that circumstance may be regarded as the probable cause (or effect) of the phenomenon.[1]

2) *The principle of double agreement.* If two or more instances in which the phenomenon occurs have only one circumstance in common, while two or more instances (in the same department of investigation) in which it does not occur, have nothing in common except the absence of that circumstance, the circumstance in which alone the two sets of instances differ is the effect, or the cause, or an indispensable part of the cause of the phenomenon.[2]

Stated in formal logical language, as these two principles are, they are not readily comprehensible. In applications of the principle of agreement, one's attention is concentrated first upon the circumstances accompanying the occurrence of some phenomenon that the investigator has chosen to study. The phenomenon to be investigated may be almost any product or process such as a variety of instances of good teaching, poor spelling, community co-operation, successful problem solving, high or low teacher morale, or good or poor financial support. The circumstances in which one would be interested are the numerous conditions that influence status. In applying this principle we seek circumstances that are common to and found among all the instances of the occurrence (or nonoccurrence) of the phenomenon under investigation.

In applying the principle of double agreement one looks not merely for factors common to the occurrence or nonoccurrence of

[1] F. W. Westaway, *Scientific Method: Its Philosophical Basis and Its Modes of Application,* Third Edition (London: Blackie and Son, Ltd., 1924), pp. 203–5.

[2] F. W. Westaway, *ibid.,* pp. 207–208.

the phenomenon under investigation but also for factors that may differentiate between the occurrence and nonoccurrence of the phenomenon. If among the circumstances accompanying the occurrence of some phenomenon there is one circumstance always present when the phenomenon occurs and always absent when it does not occur, we say that this circumstance is the cause or indispensable part of the cause of the phenomenon under investigation. The argument employed in the principle of double agreement is much stronger than that of the principle of single agreement inasmuch as we study the effects both of the presence and absence of different circumstances.

Many appraisal problems in education lend themselves to this type of reasoning. Time, money, and energy permitting, one would ordinarily prefer a much more refined procedure of investigation than that provided by the methods of agreement and double agreement. The more complex investigational procedures will be discussed in subsequent chapters of this volume. There are times, however, when it is not practicable to apply these more refined techniques and there is still need for some rough preliminary check of the phenomenon under investigation. The methods of agreement and double agreement serve this purpose. The problems studied may be almost any aspect of the school program, for example: factors in pupil morale, factors influencing the continuance of eighth-grade pupils in secondary schools, factors in adequate local financial support, likenesses and differences in good and poor achievers in algebra; likenesses and differences in the home and school background of well-adjusted and maladjusted upper elementary grade pupils.

Barr's study of good and poor teachers. Barr[1] studied the teaching performance of forty-seven good and forty-seven poor teachers of history, civics, and geography in grades seven to twelve. A fairly elaborate procedure was used in choosing the two groups of teachers:

1) A letter was addressed to a number of city and county superintendents of schools in Wisconsin, asking them to name outstandingly good or poor teachers of history, civics, and geography. Good teachers were chosen from the better-trained, better-paid teachers from cities of 4,000 population or more; the poor teachers were chosen from the less well-trained, and less well-paid teachers from smaller cities and villages.

[1] A. S. Barr, *Characteristic Differences in the Teaching Performance of Good and Poor Teachers of the Social Studies* (Bloomington, Illinois: The Public School Publishing Company, 1929), p. 127.

2) The lists of good and poor teachers thus compiled were checked against the ratings of the state inspectors. Those who did not have a rating of B+ or better were eliminated from the list of good teachers; those who did not have a rating of C— or less were eliminated from the list of poor teachers.

3) The teachers who remained on the two lists were then visited by the investigator. Teachers who did not appear to the investigator to be outstandingly good or outstandingly poor were eliminated from the study. Thus there remained a list of good teachers and a list of poor teachers each representing identical judgments of three individuals who varied widely in training, experience, and ideas of teaching. These individuals had seen each teacher at different times, with different classes, under wholly different circumstances, and agreed in classifying them as *good* or *poor*.

TABLE 22. Use of Illustrative Materials

	Number of Teachers Using Each Type of Material	
Type of Material	*Poor*	*Good*
1. Blackboard	19	30
2. Bulletin Board	0	2
3. Globe	0	0
4. Maps	16	30
5. Charts	1	14
6. Reference books	1	13
7. Map books	3	6
8. Pictures	2	11
9. Posters	0	1
10. Cartoons	0	2
11. Clippings	0	2
12. Diagrams and graphs	1	4
13. Scrapbooks	0	2
14. Lantern slides	0	1
15. Stereographs	0	1
16. Motion pictures	0	0
17. Models	0	1
18. Real Objects	0	3
19. Preserved specimens	0	0
20. Field trips	0	1
21. Demonstrations	0	0
22. Pupil experience	15	39
23. Examples and illustrations	1	5
24. Other illustrative material	0	0
25. Total number of teachers	47	47

Each teacher was visited and his work subjected to thorough analysis. The observation was guided by a carefully constructed ac-

tivity check list including such items as: teaching posture, characteristic activities, characteristic expressions, vocabulary, types of assignments, the teacher's questions, the teacher's comments, the teacher's attention to pupil responses, use of illustrative materials, economy of time, attention to physical conditions, methods of handling materials, discipline, provisions for individual differences, motivation, knowledge of the learning process, supervised study, the selection and organization of subject matter, measurement of results, characteristic pupil activities, and a number of quantitative aspects of teaching. Tables 22 and 23 summarize some of these results.

TABLE 23. Disciplinary Situations

	Number of Teachers	
Kind of Situation Observed	*Poor*	*Good*
1. Whispering	20	6
2. Giggling and laughing	6	0
3. Talking aloud	6	0
4. Foolish remarks	5	0
5. Annoying neighbor	4	0
6. Throwing chalk, paper, books, erasers, etc.	2	0
7. Walking aimlessly about room	2	0
8. Restlessness	8	1
9. Lack of attention	7	0
10. Shuffling feet, coughing	4	0
11. Talking back to teacher	0	0
12. Altercations	0	0
13. Fights	0	0
14. Total number of teachers	47	47

The author draws the following conclusions:

1) The respects in which good teachers were found to differ qualitatively from poor teachers appeared to arise from *contributing* rather than *critically* significant factors.

2) Among the more important factors contributing to teaching success are:

a) Ability to stimulate interest.
b) Wealth of comment upon pupil responses.
c) Attention to pupil responses.
d) A topical-unit-problem-project organization of subject matter.
e) Well-developed assignments.
f) Use of illustrative materials.

g) Provision for individual differences.
h) Effective methods of appraising pupil growth and achievement.
i) Freedom from disciplinary difficulties.
j) Knowledge of subject matter.
k) Knowledge of objective of education.
l) Informal conversational manner of teaching.
m) Frequent use of pupil experience.
n) Appreciative attitude on part of teacher.
o) Skill in asking questions.
p) Socialized procedures.
q) Skill in measuring results.
r) Willingness to experiment.

3) Good teachers are considerate (appreciative), patient, and pleasant, with a good sense of humor.

A number of quantitative differences were noted but the study was primarily concerned with qualitative differences.

The reading status of good and poor eleventh-grade pupils. Ankerman [1] studied the reading status of good and poor eleventh-grade pupils in English, social science, science, and mathematics attending Cass Technical High School, Detroit, Michigan. Nineteen pairs in English, eighteen in American history, twenty in chemistry, and sixteen in mathematics were established on the basis of identical ratings on the nonverbal parts of the Detroit General Aptitudes Examination and statistically significant differences on the Iowa High School Content Examinations. The testing program in reading was elaborate, consisting of the following tests:

General Reading Ability:
Progressive Reading Tests—advanced Form A:
Ability to interpret meaning.
Specific Reading Abilities:
Van Wagenen Reading scales in literature
Van Wagenen Reading scales in history
Van Wagenen Reading scales in science
Southeastern Reading test in mathematics.
Specialized Reading Skills:
Traxler High School Reading Test, Form A:
Ability to pick out main ideas.
Progressive Reading Tests, Advanced Form A:
Following Specific Directions.

[1] Robert C. Ankerman, Jr., "Differences in the reading status of good and poor eleventh-grade students," *Journal of Educational Research,* 1948, 41:498–515.

Progressive Reading Tests, Advanced Form A:
Organization of work.
Traxler High School Reading, Form A:
Rate of reading.
General Vocabulary Ability:
Seashore-Eckerson English Recognition
Vocabulary Test, Form 1.
Specific Vocabulary Abilities:
Progressive Reading Tests—Advanced Form A:
Literature vocabulary.
Progressive Reading Tests—Advanced Form A:
Social science vocabulary.
Progressive Reading Tests—Advanced Form A:
Science vocabulary.
Progressive Reading Tests—Advanced Form A:
Mathematics vocabulary.

Many tables such as the following are provided in the published report of this study indicating the statistical significance of the differences between good and poor students in the several subject areas investigated (Table 24).

TABLE 24. Significance of the Differences between the Performance of Good and Poor Students on the "Story Comprehension" Part of the Traxler High School Reading Test [1]

Number of Pairs	*Subject*	*t* [1]	*Level of Significance*
19	English	3.28	1 per cent
18	American History	3.27	1 per cent
20	Chemistry	3.28	1 per cent
16	Mathematics'	3.72	1 per cent

The author summarizes the results of his study as follows:

1) There are significant differences between the reading abilities of good and poor eleventh-grade students.

2) The reading and vocabulary abilities that differentiate good and poor achievers vary from subject to subject.

3) Good and poor students of all four fields were significantly different in their general reading ability. They were also, with the possible exception of American History, significantly different in their specific reading ability. Also, all, with the possible exception of students in mathematics, were significantly different in one and only one of the specialized reading skills each being specific to the subject field.

4) The differences on vocabulary were not consistently different.

[1] Strictly speaking the t-test is correctly applied only to randomly drawn samples.

There were other conclusions but these are illustrative of the findings of this study. Barr, in his study of good and poor teachers, was principally concerned with qualitative differences; Ankerman was principally concerned with quantitative differences.

Some precautions to be observed in the conduct of comparative causal investigations. First, it should be observed that the principles of agreement and double agreement were devised to embrace situations wherein there is a single decisive factor responsible for the occurrence and nonoccurrence of a phenomenon. Many of the products and processes of education appear not to arise, however, from a single decisive factor. Good teaching, adequate financial support, and pupil morale, for example, appear to arise not from a single critical factor but from the interaction of many factors. One may recall any number of examples from the world of physical phenomena, however, where there appears to be a single critical factor as, for example, illnesses arising from communicable diseases. There is always a seed bed, setting, or condition that sets the stage as it were, but the phenomenon under investigation appears directly traceable to a definable agent. Possibly if we were more discriminating in education we might distinguish many more types of inadequate behavior and achievement than we ordinarily do. But the point that we wish to make here is that there should be a single critical factor in the situation under investigation.

The subject of interrelationships is complex. Many persons have struggled with it. The mathematician speaks of necessary and sufficient conditions; the biologist of contributing and critical conditions; and the psychologist of field forces and vectors. Obviously, many of the problems of education arise from complex patterns of direct and secondary factors. Applications of the principles of agreement and double agreement can be regarded merely as first approximations to the truths sought.

Second, the proof provided by the principles of agreement and double agreement is not decisive. There are always many circumstances that accompany the occurrence and nonoccurrence of different phenomena. Some are incidental and some pertinent. If a number of persons who attended the same dinner and who were served the same food, became ill, and showed symptoms known to be associated with certain types of food poisoning, we may conclude that they were poisoned by some item of food eaten by all of those present. The proof is not decisive, however, the illnesses might have been due to fumes escaping from a smoking coal furnace or from a leaking gas jet. Only by supplementing these principles with those of diagnosis, discussed earlier in this chapter, can

we become reasonably certain as to the source of the difficulty under investigation.

A third precaution to which the reader's attention is called, arises out of the dichotomous nature of the phenomenon under investigation which is assumed to exist in applications of the principles of agreement and double agreement. Although we speak of the occurrence and nonoccurrence of phenomena, as of good and poor spellers, bright and dull children, successful and unsuccessful teachers, strictly speaking, these are not examples of the occurrences and nonoccurrences of a phenomenon. We seldom locate zero points on our scales, and though we may observe good spellers, bright pupils, and successful teachers, they are alike only within broad limits. Of course, this was also true of the case of food poisoning cited above, inasmuch as not all persons were equally ill. In any case, the instances of the occurrences and nonoccurrences of a phenomenon need to be defined with care.

A fourth precaution relates to the circumstances chosen for observation, study, or investigation and the readiness of the observer to see what he should. Critical factors may be present but not observed because the observer didn't think to look for them. They may have been readily observable but the mind set of the observer was such that he did not see them; or they may not have been readily observable even though there was a favorable mind set and intent to see them. In any case, the study of the circumstances accompanying any given phenomenon is a guided activity; adequate preparation must be made for such observations. The search may be inadequately guided when undertaken by novices or by experienced observers without adequate preparation. The evaluator needs to exercise very great care in the choice of circumstances to be investigated. In many instances the study of circumstances will involve complex instrumentation.

Finally in statistical research emphasis is placed upon the randomness or representativeness of samples; in applications of the principles of agreement and double agreement emphasis is placed upon variety. In the latter type of research we try to include a wide variety of instances of the occurrence or nonoccurrence of the phenomenon under investigation inasmuch as we are primarily concerned with possible exceptions. One's concern in this type of study is with universals rather than frequencies. As applied in the field of education, investigators more often than not, come out, however, not with universals but with frequencies. That is, they discover a number of circumstances that occur with different frequencies but none that always or never occurs. If frequencies are

to have meaning, however, they must arise from random samples drawn from a well-defined population and not from cases selected according to the principle of variety.

For these and other reasons it is customary to regard the method of agreement as providing a first or exploratory approach to the study of a problem and not as final proof.

The comparative-causal method provides a useful procedure for uncovering important antecedents to pupil growth and achievement. When combined with the case study method, it provides a practical pupil oriented procedure for ascertaining important interrelations among products, processes, and conditions.

Comparative-foundational studies. The second type of comparative techniques to be discussed has been principally concerned with educational theory and practice in different countries. It can, however, be equally well applied to cities, counties, states, and regions in any single country, such as the United States. Comparative education, as it is usually designated, is concerned chiefly with two questions: (1) what are the educational theories and practices of various countries? and (2) what values can be attached to them? The answer to the second of these questions is sought in social, philosophical, and historical foundations.

The comparative social-philosophical-historical foundational approach deserves wide application. In the early chapters of this volume we have emphasized the importance of well-defined educational objectives as the starting point for all properly designed appraisal projects. We would like to emphasize here the fact that school programs and educational practices must be evaluated not merely from the point of view of objectives but also from the point of view of the conditions giving rise to them. Procedures, programs, and practices are not good in general, but for various purposes, when they are appropriate to the conditions that give rise to them, and when they conform to accepted principles of goodness, rightness, or effectiveness. Much of the research relating to educational programs, procedures, and practices appears to have overlooked the importance of setting, conditions, and foundations both in research design and interpretation.

Sometimes we are concerned with the antecedents of a particular situation—that is, the attitudes developed, problems sensed, and solutions already tried. Sometimes it is the social structure of the group at any given time and place: the cultural level of the people; their mores, attitudes, and patterns of behavior; their stratification, interpersonal and group relations, and communication skills that determine the interpretation of research data.

To illustrate the comparative foundations approach to educational research, let us examine Kandel's[1] *Educational Yearbook, 1941, The End of an Era.* This volume illustrates many features of the better studies of this type. The volume is organized about topics with chapters as follows: *The End of an Era, Education and Modern Thought, Education and Politics, Administration of Education, The Education of the Child, The Education of the Adolescent, The Preparation of Teachers, Toward Reconstruction.*

In the preface to this volume the author emphasizes (1) the world setting; (2) the problems arising from World War II; (3) the pending disaster arising out of the blind acceptance of freedom and democracy in child growth as opposed to the acceptance of values and indoctrinations.

On the latter point the author writes as follows:

> This refusal to accept inherited values, to tolerate authority, to be guided by the past, had already manifested itself earlier in art, music, and literature. The era of experimentalism had already begun before educators began to question and then to discard all values and to build on a theory of creative activity and self-expression. So-called progressive education has in reality not been a new phenomenon; it came in fact at the end of a period which began with an attempt to destroy the roots of the past and ended in rootlessness. Santayana described the trend in its early beginnings when he wrote that "ideas are abandoned in favor of a mere change of feeling, without any new evidence or new arguments. We do not now refute our predecessors, we pleasantly bid them good-bye. Even if all our principles are unwittingly traditional, we do not like to bow openly to authority."

The author says at the end of chapter one that the main issue is:

> The challenge of the decades immediately preceding the war and of the war itself is to rediscover a faith which will again be based upon the recognition of the worth and dignity of human beings. The issue is not, however, solely political; the issue is to restore and cultivate a habit of mind which will not permit things to be in the saddle and ride mankind, which will strive to restore meanings and values to a world which has threatened to become, if it has not already become, meaningless. This is the supreme task in the era which opens before mankind; unless he is overwhelmed by the powers of darkness that now attack all humane values.

In his second chapter Kandel discusses such topics as: accent on change; the new education; education for a new world; the fer-

[1] I. L. Kandel, *Educational Yearbook* of the International Institute of Teachers College, Columbia University, 1941; "The End of an Era" (New York: Bureau of Publication, Teachers College, Columbia University, 1941), 393 pp.

ment of unrest; the experimental period; experimentalism in education; and the tradition of humanism.

The discussion of basic issues is continued in the succeeding chapter with treatments of the following topics: as is the state, so is education; the conflict of political theories; the totalitarian state; control of the mind; sanctions of totalitarianism; the meaning of democracy; the individual and the state; freedom and authority; the end of the state; totalitarian criticism of democracies; democracy and education in the current crises; a creed of democracy; and the crucial question.

On the last of these topics, the author writes as follows:

> The analysis which has been presented of the two conflicting theories —the monist and pluralist, the totalitarian and democratic—affects every aspect of education, in administration and organization, in curriculum content and methods of instruction, in the status of teachers, and in aims and purposes. The crucial question in the current conflict is whether the state shall dominate the minds and bodies of its subjects or whether it shall promote and encourage the growth and development of each individual citizen into a free personality with the right to think for himself and with a sense of responsibility because he is free. How this question is answered depends upon the political theory of the state and upon this theory the character of an educational system.

Following these over-all interpretive chapters Kandel discusses the administration of education, the education of the child, the education of the adolescent, and the preparation of teachers in the United States and certain foreign countries, chiefly Germany, Italy, the USSR, France, and England. These chapters are both critical and descriptive.

In the final chapter of the volume, the author discusses the following topics: all the children of all the people; physical welfare; pre-school education; compulsory attendance; articulation of schools; post-primary education; differentiation and specialization; education, a life-long process; higher education; private schools; education and the public; the teacher; education and democratic ideals.

The author concludes with the following statement about education and democracy:

> A democratic scheme of education is not complete if it is concerned only with the provision of equality of opportunity; it has also an obligation to develop a body of common traditions, loyalties, and interests as the basis of community life and stability within which the

individual can be free. Methods of free inquiry and personal responsibility acquire more meaning where this body of common traditions, loyalties, and interests already exist. The past two decades have emphasized in the theory and practice of education the growth and development of free personalities and in the end have succeeded only in producing individuals without faith because education failed to give them any meaning for life. If the present struggle between force and reason has any lesson for educators, it is that the development of personal freedom must be accompanied by the development of a sense of responsibility to and for those democratic ideals and institutions which alone can give meaning to freedom.

Kandel projects, against a rich background of scholarship and intimate knowledge of European educational development and historical foundations, a theory of education against which he criticizes current trends and practices. The study is not historical but is projected against an historical background. The study is comparative and constructively critical, providing a valuable interpretation of the current scene.

Studies such as the foregoing are concerned with broad overviews of educational theory and practice; although they may concern themselves primarily with status they do provide a means of ascertaining important interrelationships in a great variety of situations.

Some so-called comparative-foundational studies are neither comparative nor foundational. Many studies superficially labeled comparative are not comparative; neither do they involve evaluations made in the light of foundations. A good historical study of educational developments in almost any country, state, or region can be immensely worth while but it is not a comparative study until comparisons are made between the practices of two or more states, countries, or regions. Those conducting studies of the comparative foundations type are strongly urged to follow a two-step process: (a) the comparison of practices to discover likenesses and differences; and (b) the appraisal of practices in the light of foundations. The historical type of study will be considered in a later section of this chapter.

Practices are not easily categorized. Even when the languages of two or more countries for which practices are being ascertained and compared is the same, the categorizing of practices is difficult; but when the languages are different the task becomes almost impossible. The concepts and language habits of peoples of the same language as (for example, England, Australia, and the United

States) may be nevertheless very different, while those with different experiences, concepts, and languages become almost impossible to comprehend except with long and varied contacts with the peoples concerned. The types of information gleaned from amateur translations or short visits are likely to be more harmful than helpful. The coverage even by symbols that would appear identical varies from language to language and people to people. To categorize practices under such conditions is a most difficult process and likely to be misleading.

Statements about the prevalence of practices in various countries must be made with care. One notes many loose statements in the literature of education about educational practice. When can a practice be said to be characteristic of a state, country, or region? What constitutes prevalency? Possibly a practice is characteristic of or prevalent for only some segment of the population or some geographic division. Only through carefully limited statements based upon fact can one hope to arrive at dependable statements about educational practice.

Comparative-foundational studies may be undertaken as complete and separate entities in and of themselves; or as parts of other descriptions and appraisals of status. A more general use of foundations materials as an integral part of all appraisal studies would probably prove valuable to all concerned. Both types of comparative studies, namely, the comparative-causal and the comparative-foundational approach, may be effectively applied to or carried forward concurrently with the historical type of study to be discussed next.

HISTORICAL STUDIES

Many of the studies already referred to in this chapter have their historical aspects, as for example, case histories and foundational studies of educational theories and practices. Over and above this, all educational problems and practices have their historical antecedents. The school elders of any state, county, region, or nation, might well, if given an opportunity to say so, tell us that this and that problem and practice has a history—and a history that cannot be ignored if we desire the fuller understanding of why things are as they are and if leadership is to be sensitive to the complex social forces that make for progress. Whether one turns to a review of previous investigations as one does in initiating a scientific investigation or to recollections of elders as referred to above or to

records of events and transactions as one does in the study of court records, legislative enactments, old newspapers, and correspondence, one turns to history for information and guidance.

History may be defined broadly as any appeal to past experience for help in knowing what to do in the present and future. It may be found in the memories of men but more generally in *documents*—a technical word used by historians to include a variety of sources of information relative to the past: *records,* such as legislative acts, official papers, newspapers, periodicals, autobiographies, memoirs, annals, charters, diaries, letters, maps, court decisions, photographs, films, recordings, paintings; and *remains,* such as monuments, buildings, tools, utensils, weapons, and clothing.

The sources of information may be primary and secondary. The reference above was to primary sources; but it is still history, even when one utilizes the accounts of others. Most of the histories of education published throughout the years fall within this category. History is concerned with the sum total of past experience: social, political, vocational, and intellectual—wherever documents are available to tell us what this experience was. Much of what is educational will find its meaning in social and cultural foundations.

The educational historian is concerned with two types of problem: (1) What are the facts? and (2) What significance have they for the solution of educational problems? Let us begin by considering some of the problems associated with ascertaining the facts.

The search for sources of information. Having selected a problem for investigation, the investigator must seek sources of information. The *Educational Index,* the library card catalogue, and secondary sources will ordinarily be those first consulted. But the preparation of a scientific report involves much more than even a careful reading of secondary sources. In serious research one will search for documentary evidence and original sources. Among the aids available to the student are bibliographies of various kinds, encyclopedias, source collections, and numerous periodicals. Data may be found in court records, legislative enactments, diaries, newspapers, memoirs, private correspondence, and other contemporary records. Beginning with indexes, bibliographies, and general references one works deeper and deeper into his subject; each source is scanned for evidence of yet other sources until every possible available source of information has been examined. A complete survey of the original materials may require visits to many libraries, public and private files, local, state, national, and international, depending upon the scope and importance of the topic.

Determination of the facts. Following or accompanying the search for materials, one attempts to ascertain the facts with reference to the subject under investigation. The determination of the facts is not always easy. Trained historians proceed with great care. Through a process of external and internal criticism the historian establishes the accuracy of each document. External criticism is concerned with the genuineness of the document itself, whether it is what it purports or seems to be. Internal criticism is concerned with the meaning and accuracy of the statements within the document, after spurious and interpolated materials have been removed from it. Although in practice there may be no sharp line of demarcation between these two phases of historical criticism, inasmuch as both may proceed simultaneously or with a great deal of overlapping, it is helpful for purposes of discussion to treat each separately.

External criticism. The first test which the historian applies to a document or remain is that of its genuineness.[1] The search in this respect involves textual criticism and investigation of authorship. There may be errors in the reproduction of documents, printers' errors and copying errors, intentional and unintentional, that lead to "corrupt" or "unsound" documents. The opportunities for error are numerous. In establishing the genuineness of a document, it is important to know where it comes from, its authorship, and date of publication. In the case of ancient documents and remains the process is elaborate.

Internal criticism. Internal criticism involves the mental operations which begin with the observation of a fact and end with the written statement. It is concerned with two types of operation: [2] (1) the analysis of the content of a document to ascertain what the author meant; and (2) the analysis of the conditions under which the document was produced to verify the author's statements. Among the reasons for doubting the good faith of an author are (1) the author's interest; (2) forces of circumstances; (3) sympathy or antipathy; (4) vanity; (5) deference to public opinion; and (6) literary distortion. Some of the reasons for doubting accuracy are (1) the author was a poor observer; (2) hallucinations, illusions, and prejudices; (3) negligence and indifference; and (4) the facts may not be of a nature to be directly observable. The determination of particular facts involves a very careful sifting of evidence, since they may have been observed and recorded under

[1] C. V. Langlois and C. Seignobos, *Introduction to the Study of History* (New York: Henry Holt Company, 1898), 350 pp.
[2] *Ibid.*

conditions far from ideal. When one recalls the care exercised in experimental research in defining terms, collecting, recording, and analyzing data, in developing controls, and in generalizing, one becomes acutely aware of the limitation of historical research. In any case, one has the problem of taking the evidence such as it is and reconstructing the past to the best of one's ability. Out of operations such as the above one ascertains the facts.

Interpretation of the facts. Facts are not ends in themselves but the materials for intellectual contemplation. In historical research one comes to the facts with questions to be answered. One may wish answers to questions such as the following:

1) What are the major characteristics of some phase of school education in some particular time and place?

2) What is the sequence of events leading to some problem situation that may be under investigation?

3) What circumstances appeared to be potent in bringing about the more important changes that have taken place in the historical background of some problem situation?

4) What are the trends for some particular period of history, country, or phase of educational development?

Ascertaining the major characteristics of some phase of school education for some time in the past. Such studies may be concerned with outcomes, processes, or conditions and their interrelationships insofar as we can discover them from the evidence found in documents and remains. They may relate to selected phases of the school program or to the over-all pattern. They may be made for any period of time or for any country. Insofar as they are limited to ascertaining the facts about some phase of school education for a particular past time and place and nothing more, they are status studies, such as those, discussed in an earlier chapter, and subject to all the restrictions placed on such studies. The major difference between the type of study discussed here and that discussed in the chapter on the description and appraisal of status is that in the historical studies we depend upon records and remains for our evidence, rather than upon living subjects. The documents and remains to which we go for evidence may contain the data that we seek in records of contemporary surveys; if not, the time has passed when additional data may be collected. In many instances, we are concerned, however, not merely with status for some remote time and place but with the social foundations and sequences of events that lead to things in the present. This is a very much more complex operation and will be discussed shortly.

In seeking answers to questions about the status of things in the

past, one must exercise many precautions if the true state of affairs is to be ascertained. As in surveys of the present, there are many opportunities for error, both in ascertaining and interpreting the facts. One of the very common sources of error will be found however in statements about the frequency with which various events or characteristics occur or the extent or their occurrence among the total population, subgroups thereof or for different geographic areas. Such generalizations will be made only on the basis of verifiable statements found in the records. These statements may be about individual instances on the basis of which the investigator formulates a generalization; or the generalization may be found ready-made in the records. In any case, statements about the frequency with which various phenomena occur and their extent must be made with great care.

The sequence of events leading to some problem-situation. Another type of question to which the educational historian seeks an answer is that involved in tracing the sequence of events leading to some problem situation under investigation. Such information should throw light upon prior courses of action, their effectiveness, and the attitudes of people toward them. Effectiveness is, however, only one of the many criteria that may be applied to educational events. Many reasonably effective courses of action are not desirable for the simple reason that the people do not desire them. In an absolute sense, one course of action may be very much better than another, but the experience of the persons involved may be such that they prefer the less effective. The original test of effectiveness may have involved a very limited concept of effectiveness; or the persons involved may merely have had an unfortunate experience or through spurious reasoning or ungrounded fear they may have come to prefer a less desirable course of action. In any case a thorough knowledge of the sequence of events leading to the problem under investigation will usually throw new and valuable light upon the problem at hand.

Studies such as those suggested above involve an intimate knowledge of the attitudes and motives of people and of cause-and-effect relationships. Investigators of attitudes and motives will need besides a thorough grounding in individual and social psychology considerable practical insight into human nature. In addition to the individual and group dynamics of the situations there will also be many static factors, such as those found in custom and tradition and the socio-physical aspects of the situation, past and present. Obviously generalizations about these will need to be made with care.

Studies of the potencies of various historical antecedents. The problem here is similar to that of any other scientific investigation except the data are to be found not in new appeals to experience but in documents and relics. There are two types of situations for which one seeks such information: (1) for individual events definitely located at some particular time and place; and (2) for similar events occurring in a variety of times and places. The first type of study obviously precedes the second and supplies the data for it. From a summary of the data for a particular case one comes to statements about the potency of various factors in this event, happening and situation.

The data may consist of frequently stated urgencies and attitudes or contemporary summaries of the situation, found in the records. In such instances a faithful description of some past happening or situation will be the goal of the research worker. In the second instance, the investigator's concern is more inclusive. He may, for example, seek some more inclusive generalization true for a variety of situations that have occurred in remote times and places. We seek through applications of the logical principles of agreement and double agreement to profit from history.

As in all other types of research, appeals to history for answers to questions such as the above are accompanied by many difficulties. In establishing the facts for individual happenings and events one performs operations similar to those undertaken in making case studies, except in the latter instance the subject is alive and available for further observation. It will be recalled that it has already been said that case studies may be made for communities and institutions as well as for persons. In deriving generalizations about a variety of situations, one employs operations already discussed in the immediately preceding section of this chapter on comparative studies. The principal logical tool is that of the principles of agreement and double agreement. History being what it is, it is only natural, however, that many of its value statements will be based upon social foundations. There will be many comparative education studies of past times and events.

Studies of trends. The determination of educational trends involves a careful synthesis of materials from many sources. The purpose of the *trends study* is to establish a course of events originating in the past that may be expected to continue in the future. From examination of the facts for some particular period of history a generalization is reached about things to come. The process is complex and fraught with many pitfalls. In addition to establishing the facts for certain particular periods of time, the investigator makes predictions about purposes, persons, principles, and conditions, as

well as courses of action to come. Human nature being what it is, values continuing as they are, and conditions being similar to those already investigated, such and such events may be expected to take place.

Trend studies may be undertaken with reference to conditions and values as well as courses of actions. Crawford [1] makes the following suggestions for trends studies: (1) a trend can only be measured by a comparison between conditions or practices at different times; (2) it is preferable to study trends at long intervals rather than short intervals; (3) it is better to compare and contrast separate time-periods than adjacent ones; (4) the conditions or practices should be classified on an identical basis; (5) the reduction of data to quantitative form tends to decrease the subjective errors commonly found in verbal descriptions; (6) changes in curricular material may be determined by an analysis of textbooks over a period of time; (7) analyses of professional magazines will show new emphasis and trends for any period of time; (8) publication dates of textbooks can be used to ascertain the rate of increase or decrease of the number of books with respect to the various school subjects; (9) the number of articles reported in periodicals on a certain topic for a certain time-period will indicate its rise or fall in popularity as a subject of discussion; (10) a trend may be ascertained by noting the number of entries in the *Education Index* or other periodical indices; (11) old examination questions or questionnaires may be readministered and the results compared with those secured from some earlier time; (12) the research worker should be very careful to choose materials which are truly representative of their respective periods; and (13) the existence of a trend should not be taken as its justification.

Special problems and areas of application. The purpose of the foregoing discussion has been to present a brief summary of what is involved in historical research to orient the student with reference to it and show its relationship to other types of educational research and appraisal. There are many excellent detailed discussions of the methods of historical research such as those provided by Langlois and Seignobos,[2] Hackett,[3] Edwards,[4] Brickman,[5] and

[1] C. C. Crawford, *The Technique of Research in Education* (Los Angeles: University of Southern California, 1928), pp. 58–60.

[2] C. V. Langlois and C. Seignobos, *Introduction to the Study of History* (New York: Henry Holt and Company, 1898), 350 pp.

[3] H. C. Hackett, *Introduction to Research in American History* (New York: The Macmillan Company, 1931).

[4] Newton Edwards, *The Courts and the Public Schools* (Chicago: University of Chicago Press, 1933), 582 pp.

[5] William W. Brickman, *Guide to Research in Educational History* (New York: New York University Bookstore, 1949), 220 pp.

Woody[1] to which the student about to engage in systematic work in this field is referred for guidance and assistance.

As one turns to these more detailed guides, attention should be called again to the unique place and function of history in the family of research techniques. (1) It should provide, when carefully pursued, a helpful overview of the published researches and previous experiences of other persons and groups of persons with the problem under investigation. From an historical point of view these background chapters are frequently poorly done. (2) It should provide a valuable frame of reference for the further evaluation of all facts and generalization whatever their source. Such facts and generalizations may arise from excellent experimental, statistical, and clinical studies, but their full meaning cannot be had except from a very careful scrutiny of their social and historical foundations. (3) It provides a valuable method of research in and of itself, especially when undertaken in conjunction with the comparative methods of research. Everything has a history; students of professional education would profit from a better knowledge of it.

SUMMARY

In a previous chapter it has already been pointed out that in appraising status one may appraise products, processes, and conditions. There are, however, many complex relationships within and among the products, processes, and conditions. The purpose of this chapter has been to discuss some of the methods that one may employ in exploring some of these interrelationships. In this chapter we have discussed some of the ways that employ nonmathematical techniques, such as the case study, comparative study, and historical study. Some of these techniques employ elementary statistical devices but are largely nonmathematical in character.

Appraisals of products, processes, and conditions should be made with reference to the persons involved. Sooner or later the educative processes must get down to the individual pupil. His particular qualities, considered in relation to the social, physical, and biological setting in which they are found will determine his needs and ultimately the objectives of education. The study of qualities provides important data and points of view from which pupil growth and achievement are appraised. Whether a particular product, process, or condition is satisfactory will depend upon the persons involved. The case study method provides one method for as-

[1] Thomas Woody, "Of history and its method," *Journal of Experimental Education,* 1947, 15:175–201.

sessing some of the personal factors in pupil growth and achievement.

Appraisals of the products, processes, and conditions of education must also be made with reference to the social, physical, and biological setting for learning. Pupil behavior does not take place in a single, uniform, and constant environment. From a study of comparative education one learns about likenesses and differences in the practices of various cities, communities, states, regions, nations, and how these practices have arisen out of (and find their justification in) social foundations. History, when it includes social and intellectual history as well as political history, provides a valuable frame of reference and point of view from which appraisals should be made.

CHAPTER VIII

Experimental Design

The role of experimental evaluation. Many factors limit or facilitate pupil growth and achievement. Some of these factors are within the control of the school; others are not. To the extent that the conditions favorable or unfavorable to pupil growth and achievement can be ascertained and controlled, to that degree may those responsible for school education facilitate the learning-developmental process. The intelligent modification of current educational practice can come only from a better understanding of the biological, psychological, and environmental factors that influence learning outcomes. Educationalists need precise information concerning desirable and undesirable effects of the many factors that operate to affect pupil behavior. The way to secure this information is through a program of continuous experimental evaluation.

This, then, in broad outline is the aim of experimental evaluation: to discover the laws and underlying conditions basic to the growth of pupils, and to determine the ways in which the school can promote efficient development. The principal means employed for this discovery is the experimental method; that is, observations under conditions which the investigator can control.

The experimental method provides the means of establishing a valid basis for drawing inductive inferences. Here, as in all the sciences, new observations form the basis of new knowledge. In order to establish the same secure basis that science demands, however, it is essential that the observations be obtained under conditions which will make rigorous inferences possible. This requires that certain principles of planning, executing, and interpreting an experiment be applied. Experimental observations are experiences

planned in advance in accordance with the principles that will make an unambiguous interpretation of the observational data possible.

Principles underlying the design of experiments. Evolution of the principles of experimental design is closely connected with the concept of precision. By precision is meant the amount of information relevant to the hypothesis under test that it is possible to make an experiment yield. The theory of experimental design is intended to supply improved techniques, efficient methods of arrangement, and careful analysis of experimental results. The object of selecting an arrangement, like that of any good technique, is to secure the greatest accuracy commensurate with the amount of funds, time, and labor available for carrying out the experiment. Developing a proper experimental design requires careful planning, including consideration of possible results and their interpretation. One of the commonest defects in experimental design is failure to anticipate and thus to establish safeguards to prevent biased results.

A great part of scientific activity consists in constructing a theoretical system in order to represent the facts as accurately as possible. Theories are useful in stimulating experiments which are designed to test them and which may in turn lead to modification and reformulation of the theories themselves. To this extent any experiment is related to previous experiments and to previous experiences. But regardless of the complexity of the problem under investigation, the aim of the experimental design should be to make the experiment self-contained. That is, the experiment should be capable of providing its own evidence as the basis for a valid and unambiguous interpretation, making it unnecessary to appeal to other experiments or to previously acquired experience.

A primary requisite of a self-contained experiment is the provision of controls, which enable the experimenter to base his conclusions on the differences observed in the reaction of similar groups of experimental subjects who have been subjected to some accurately defined difference in treatment. The term "treatment" is used quite generally to designate the factor or factors the effect of which the experiment is designed to test. Adequate controls serve to exclude causes other than those under evaluation which may have produced the result observed in the experiment. This the control does by enabling the calculation of the probability that the causes could be other than those assigned. In the case of a

psycho-physical experiment, for instance, where the experimental results consist of "right" or "wrong" judgments, the level of probability is capable of computation on the basis of the *number* of subjects constituting the control. In an experiment where the results are given in quantitative form, the computation of the level of probability is based on the *theory of errors*.[1] For example, in an experiment designed to compare the relative efficiency of the non-oral and the silent and oral methods of teaching reading, the experimental results could be measured by scores on a standardized reading test. In this way, the experimental evidence can be made objective in that it can be made independent of the varying values different interpreters might attach to changes in the experimental subjects if no controls were involved. Through use of controls results of experiments become comparative.

A second requisite of a self-contained experiment is that the experiment must be planned in such a manner as to secure a valid estimate of the experimental errors by which the comparisons made are affected. The variation in the behavior of the experimental subjects unexplained by the experimental controls is spoken of as due to experimental errors. The experimental plan should permit a valid estimate of experimental error as well as make possible an unbiased comparison between the treatments under test. The estimate of experimental error is essential inasmuch as it furnishes the basis for the application of the test of significance. A test of significance is the technical term given to the procedure by which a critical examination is made of the reality of the differences observed from contrasted treatments. In addition to obtaining the best or most efficient estimate of the differences between treatments presumed to exist, it is essential to establish whether the presumed differences are real or attributable to random variation. The basis for this determination is provided by the test of significance.

It should be clear from the discussion above that the application of statistical methods is not restricted to the analysis of observational-experimental data but plays an essential part in the planning of experiments. Thus, experimental design and statistical analysis are not independent problems. Both are aspects of the same problem. Both provide a secure basis for making possible new additions to knowledge.

[1] Here the basis of the probability grows out of the number of independent comparisons among the observational values available for calculating the estimated error.

DESIGN, EXECUTION, AND INTERPRETATION OF THE SCIENTIFIC EXPERIMENT

Fundamental problems in the design, execution, and interpretation of an experiment will be discussed under the following headings:

1) Origin and definition of the problem
2) Formulation of the hypothesis
3) Specification of the populations sampled
4) Grouping and pairing to secure homogeneity
5) Control of nonexperimental factors
6) Selection and measurement of the criterion
7) Analysis and interpretation of the experimental observations.

Origin and definition of the problem. Educational research bristles with problems to which the experimental method can be applied. These problems grow out of critical observation of actual educational conditions. They are encountered in actual contact with the conditions as they exist. Inquiry arises out of the intellectual desire to interpret the conditions observed. Interest springs from the desire to understand the underlying processes and to gain control over them. As a result educational practices may be improved.

An active classroom environment is a complex situation. Different viewpoints, growing out of both practical and theoretical considerations, exist with respect to the efficacy of different procedures. The educator is interested in the discovery and explanation of optimum conditions for learning. In seeking scientific explanation the sufficient conditions for the appearance of phenomena will ordinarily be sought out first; then the quest for the necessary conditions begins. The plurality of alternatives serves to make investigations extensive and flexible. Experimentation discloses new conditions which in turn create new problematic situations. The development of a critical attitude followed by analytical observation and systematic comparison of educational practices provides the intellectual stimulus and challenge to continuous experimentation.

Beginning with a felt need or difficulty, one must eventually describe the regions of inadequacy more precisely. A rigorous and accurate definition of the research problem is the requisite initial step. Definition of the problem should contain the distinguishing

characteristics of the two or more sets of conditions under comparison. The postulated outcomes of the respective treatments should be stated in operational terms. When the experimental subjects are human beings, the outcomes should be stated in terms of behavior. The statement of experimental outcomes is meaningful only if it can be tested; hence it must be amenable to observation and measurement. If two educational procedures have different purposes, no valid comparison is possible. If, however, there are distinctive outcomes for either procedure in addition to those that are common, all should come within the scope of the evaluation program. The definition of the problem should also make definite and precise the status of the non-experimental factors, the length of the experimental period, and the specification of the population from which the experimental subjects were taken. These topics will receive further consideration later in our discussion.

Formulation of the hypothesis. The ability to formulate hypotheses has greatly facilitated progress in science. There are no golden rules by the application of which hypotheses originate. Knowledge of the field gained through experience and experimentation however is essential to the formulation of promising hypotheses. A number of hypotheses are usually suggested and entertained; the research worker must select from plausible hypotheses those which are reasonable and promising. Working hypotheses guide the experimenter in planning his inquiry, and instigate and direct the data to be collected and used. They also disclose the conditions under which the evidence will have the maximum influence in relation to the decisions to be made. Often several reasonable hypotheses may be in agreement with the data. Fresh data need to be collected before differentiation among hypotheses can be made. In the earlier days of science, an experiment leading to such discrimination was called the "crucial experiment."

Even without the collection of new data, one hypothesis may be accepted in preference to the other. One hypothesis which a scientist favors in the examination of his data is that chance is the cause of the effect or that the observed variations and effects are attributable to random factors giving rise to random errors rather than that they are the resultant of newly discovered causes. This kind of hypothesis lends itself to statistical evaluation more readily than the positive assertions commonly employed in non-technical discussions. The preference for this hypothesis derives from a canon of science referred to as parsimony. This principle suggests use of the simplest explanation of the facts and the one which introduces the least number of new quantities, constructs, or ideas.

The aim of science is to explain a maximum number of facts with a minimum number of concepts.

The hypothesis to be tested by an experiment must be rigorous and exact; it must have testability. It is a requisite of science that its content must be capable of being refuted if it is to have a scientific meaning. This sort of hypothesis has come to be known as the *null hypothesis.* In this form, an hypothesis is never confirmed, but it may be rejected. An experiment permits the facts of observation to refute the null hypothesis. In the language of the experimenter, the usual form of statement of the null hypothesis is that there is no difference in the effects of the treatments under comparison except those arising from sampling or other chance factors. For example, in an experiment involving a comparison of two teaching methods, one form of statement of the null hypothesis would be: there is no difference between the two methods with respect to the postulated outcomes. It could also be stated thus: the outcomes are the same for the two methods of teaching. In the language of the statistician, the null hypothesis could be stated as follows: the true difference between the two means of the groups is zero, or that the two samples are from the same population.

Specification of the population sampled. In planning an experiment, specification must be made of the population for which a conclusion is to be drawn. The experimental results are obtained for a specific group of individuals called a sample. The conclusions to be drawn are for the population of which the sample is usually a very small part. Thus it may be said that the particular experiment is of interest in itself only insofar as it provides information basic to the formulation of conclusions about the population. It is the investigation of the constant relations between events that constitutes a scientific problem. Furthermore the relations sought in science are those which can be generalized. This means that inferences may be drawn from experimental results in a single experiment or series of experiments which apply to all cases of a similar kind. Thus the subject matter of science is not restricted to particular time-and-place situations.

The concept of an infinite parent population is fundamental in statistical inference. Although the idea of an infinite parent population is a mathematical abstraction, it is basic in the interpretation of the results of experiments. Thus we can conceive of any finite sequence of repetitions of a random experiment as a sample from the hypothetical infinite population of all experiments that might have been carried out under the same conditions. Experiments

even if replicated with the utmost care to maintain constant conditions, yield varying results. Experiments of this kind may be said to be random. Interpreting a random experiment in terms of sampling assumes that the experiment is so arranged that the probability of being chosen is the same for all members of the infinite population of experiments. The experimenter is interested in estimating the "true" value; that is, the value corresponding to the frequency of the experimental event—say E, in the total infinite population. It is a logical assumption that for indefinitely increasing values of N (the number of individuals in a sample), the frequency of the event, E, would ultimately reach the "true" or population value.

In dealing with observations on individuals from human or other biological groups, the populations studied are actually finite. The statistical model based on an infinite population can be conceived as the limiting case of a finite population where N, the number of individuals, is assumed to increase indefinitely.

The idea of a random experiment, as one from an infinite population of experiments carried out under uniformly constant conditions, involves the designation of the experimental subjects as an essential part of the uniformity of conditions. The samples of individuals constituting the experimental subjects must then be representative of the population for which the conclusions from the observed experimental results are to be drawn. Thus in an educational experiment, the experimenter must be able to show that the students involved in the experiment are representative with respect to the characteristics (for example, learning characteristics in an experiment dealing with learning) of the universe of students of which they constitute a sample. Unless the experimenter can provide evidence necessary to show that his groups are representative, he should refrain from generalization and limit his interpretation to the specific conditions of his investigation. He can construct a hypothetical universe for which he might generalize, but this is usually of little practical significance. It might be doubtful in this case that he has conducted an experiment, in the scientific sense.

The attempts made to select groups of individuals so that the composition of the groups under experimentation may be the same or comparable involves, in last analysis, the problem of securing representative samples. Some of the devices that have been designed to secure homogeneity of groups of individuals for experimental purposes will be discussed.

Grouping and pairing to secure homogeneity. The chief concern in experimental work is the necessity of securing a sample which may be regarded as representative of the population from which it is drawn. Certain precautionary steps must be followed in constructing a sample in order that it will be effectively representative. Thus precaution must be taken to make a sample such that there has been no bias in selection in regard to any characteristic related to the experiment. For example, in an experiment comparing oral and nonoral methods of teaching reading, it would be desirable to take at least two samples of children, one to be taught by the nonoral method (the experimental group) and the other by the conventional method of silent and oral reading (the control group). For a valid test of the relative efficacy of the two methods, it would be essential that the method of securing the experimental and control groups does not introduce any bias towards greater achievement of one group over another. There may be differences in intelligence, chronological age, reading readiness, school grade, and many other educational factors related to the criterion.

The central problem here is the choice of means whereby experimental precision may be increased. It is hoped that groups may be selected so as to be as homogeneous as possible at the beginning. It is essential that comparisons be made under conditions made as equal as possible in all respects except the factor to be tested. This idealized condition, aiming at perfect experimental control, is capable of being carried out only with varying degrees of approximation. With increasing knowledge of the principles of experimentation, the successive approximations may be expected to become more and more accurate.

The principal means for securing equality of experimental materials for comparisons and contrasts of experimental treatments are (1) random control grouping, (2) paired control grouping, (3) technique of co-twin control, (4) statistical controls.

The random control method consists in drawing individuals at random from the same population to constitute the experimental and control groups. This method affords a basis for avoiding bias in that the differences between the groups in respect to characteristics, such as those mentioned above, will most likely not exceed a chance amount which is readily calculable. With large numbers of subjects the various characteristics may be very nearly equalized; with smaller numbers such equalization will be somewhat less exact. A variant of the random method is sometimes fol-

lowed in arrangement of two groups so that they will be approximately equal with respect to the means and standard deviations of the measured variables.

This method of "random control grouping" may be made more exact if "paired control grouping" or "equivalent matched grouping" is substituted. Here measurements are made on the individual experimental subjects on certain basic characters that bear a relationship to the experimental criterion. Individuals are then paired off on one or more of these characters so that the composition of the experimental and control groups respectively is as nearly homogeneous as possible. At times qualitative characters may also be used in pairing. Intelligence as measured by group or individual intelligence tests, previous achievement, initial status with respect to an experimental criterion or criteria, special aptitude tests, sex, socio-economic status are examples of basic characters ordinarily considered. Since the number of subjects available for the experiment is often limited, it is difficult, if not impossible, to secure two or more groups of individuals equivalent with respect to all basic characters. The method sometimes used is that of equating individuals as closely as possible with respect to one character, e.g., intelligence test scores. After this has been done the members of a pair are then randomly assigned to either the experimental or the controlled group. When as many individuals as desired have been assigned to these alternative categories, on the basis of the single character, statistical measures such as the means and standard deviations can be calculated for the other basic characters, which are quantitatively measurable as a further check on comparability. In case of significant disparities in any of the traits, individuals may be shifted to ensure homogeneity. Other procedures are employed at times. In some cases the individuals are paired after the experiment has been completed. In small schools where there is only a single class of pupils for a given year, the two classes in successive years may constitute the experimental and control groups. The critical research worker is aware, however, that individuals paired on one or more control factors does not mean that the individuals are identical. Randomization and replication are necessary in order to estimate the closeness of the matching.

The utmost uniformity in experimental material is attainable when identical twins are used as experimental subjects. This technique of "co-twin control" is used effectively in a study of development and maturation. Gesell and Thompson, for example, used two identical twin girls to study the degree of developmental simi-

larities and of developmental divergence that might result from specific training given to the one but not to the other twin.[1] The differential effect of contrasted educational treatments might be studied profitably by use not only of identical twins as experimental subjects but also of identical twins as teachers. Obviously, the dearth of identical twins precludes wide use of this technique. It serves however to illustrate the desirability of controlling variation by planning the experiment in order to apply contrasting factors upon experimental subjects as homogeneous as possible at the beginning.

Studies conducted to measure the relative importance of hereditary and nonhereditary or environmental factors have also used methods involving contrasting groups such as the following:

1) Comparison of identical twins reared in different environments with children of different parentage reared in the same environment.

2) Contrast between the different categories of twins (identical and nonidentical), and between them and ordinary siblings and unrelated children.

3) Comparison-contrasts between foster children reared in different homes as well as between foster children and their brothers, between foster children and their foster brothers, and between foster children and their own and foster parents.

Although the process of pairing serves to increase the sensitivity of the experimental observations, a considerable practical difficulty is encountered in obtaining representation throughout the range of variation. There is difficulty in securing a basic character that correlates highly enough with the criterion to ensure closely homogeneous groups. There is also considerable loss in information in that couples cannot be found for all individuals particularly if the standards of pairing are rigorous. In any case, selection is always a hazardous process since the nature of the sample becomes more uncertain as more or less arbitrary selection of individuals to be paired takes place. A further difficulty arises from the fact that the difference in the mean effect of the contrasted factors may depend upon the character or characters used in matching.

Because of the difficulties and limitations of these commonly employed methods, increasing use is being made of statistical controls. Through application of the technique of analysis of covariance the necessity of matching individuals disappears and hence all indi-

[1] A. Gesell and H. Thompson, "Learning and growth in identical twins," *Genetic Psychological Monographs,* 1929, 6:1–124.

viduals can be used. The process results in adjustment in the means of the contrasted groups for whatever inequalities exist in the basic characters of matching. Thus the evidence provided by the data themselves is the source of corrections for inequalities. In this sense the experiment may be said to be self contained.[1]

A further method for overcoming difficulties in securing comparability is that known as the Johnson-Neyman technique. This technique makes possible the specification of a population in terms of the basic characters of matching. When the null hypothesis under test is rejected a region of significance is set up. From the properties of this region, the experimenter can specify for what kinds of students as described by the matching variables a difference in the criterion exists. It is also possible to discover the situation sometimes resulting from contrasting treatments where one method, or treatment, is superior for one type of student and another method superior for another type of student.[2] For example, supervised study may be desirable for less able pupils, but undesirable for superior ones.

Control of nonexperimental factors. The search for causes in experimental science is guided by the principle of causality. Qualitatively this principle may be stated in the proposition, "the same cause produces the same effect." It is in the invariant relation of occurrence between two or more changes of events that the concept of causation is identified. Causality is identical with predictability. The aim of experimentation is to determine the sufficient conditions for the causes which produce the scientific phenomenon; it is also concerned with establishing the necessary conditions. As illustrated in the physical sciences, the control of the conditions under which a causal law operates is achieved by successive approximations. In the preliminary stages of experimentation an approximate law is obtained by assuming that the conditions are uniformly constant. With this formulation it becomes possible to define the conditions of the experiment more exactly and thus

[1] For illustrations of the process of covariance analyses see William C. Cochran and Gertrude M. Cox, *Experimental Designs* (New York: John Wiley & Sons, 1950); Max D. Engelhart, "Suggestions with respect to experimentation under school conditions," *Journal of Experimental Education,* 1946, 14:225–244; Max D. Engelhart, "The analysis of variance and covariance techniques in relation to the conventional formulas for the standard error of a difference," *Psychometrika,* 1941, 6:221–233; Palmer O. Johnson, *Statistical Methods in Research* (New York: Prentice-Hall, pp. 75–80, 276–326); E. F. Lindquist, *Statistical Analysis in Educational Research* (Boston: Houghton Mifflin Company, 1940), pp. 76–86.

[2] For illustrations of this process see Walter L. Deemer and P. J. Rulon, *An Experimental Comparison of Two Shorthand Systems* (Cambridge, Mass.: Harvard University, 1942).

reach a closer approximation. This prototype of experimentation is essentially that followed in the experimental biological sciences, although control of conditions is usually a much more difficult problem than in the physical sciences. The method consists in producing the circumstances surrounding a phenomenon in several ways and combinations, constantly under experimental control. Skill in this method resides in the appropriateness of the selection of combination of circumstances and in ingenuity in devising the required control. The problem of controlling the nonexperimental factors is especially difficult.

Nonexperimental factors are those conditions in the system isolated for experimentation, which should be uniform for both the experimental and control groups. It is an essential aspect of experimentation to obtain evidence which will make possible the determination of the extent to which uniformity of nonexperimental factors has prevailed. Obviously, if factors other than the experimental factor have been disparate during the course of the experiment, it will be impossible to obtain a valid estimate of the effect of the experimental factor. If allowances can be made for the effect of lack of uniformity of nonexperimental factors, an approximation may be made of the effect of the experimental factor, but this provides a very hazardous basis for interpretation.

If we take, for illustration, the simplest case of all, a comparison of two "treatments," it is not sufficient to arrange two equated groups assigned respectively to the treatments to be tested, and argue as to the results on the basis of the observable characters established as the experimental criterion. Common sense dictates that the two groups shall be identical with respect to the potential achievement of the criterion, and that they shall be treated alike in all other respects except for the factor to be tested.

Such complete experimental control is an ideal, never capable of being fully achieved. The equality of groups might be tested in advance of the experiment by applying the same treatment and evidence as to equality obtained. But the measurements of equality are subject to error. The criterion measures themselves are subject to inevitable errors. The numerous nonexperimental factors cannot be made identical for the two experimental groups. Even after all variables which might conceivably influence the results have been listed and "equated" there are others not suspected. For example, in a learning experiment on techniques of instruction, relative skill in using techniques, zeal of the teacher with respect to an experimental factor, materials of instruction, the time allotted to the learning activity, unbiased measurement, all are examples

of factors that must be considered in securing uniformity of conditions. Consequently it is not sufficient to demand that all conditions be exactly alike in every respect except the factor to be tested. This requirement is never possible in any kind of experimentation.

The equalization of conditions, other than those under investigation can, therefore, be achieved only to a greater or lesser degree. This holds true no matter what amount of experimental skill and attention may be expended in designing an experiment. The essential step in the experimental procedure which serves as a safeguard for valid interpretation of experimental results is, therefore, that of randomization. After all precautions have been taken to secure equality of conditions, the individuals constituting the experimental subjects are by this method assigned by a random process to one or another of the treatments under experimental test.

Even when all these precautions are observed, it is important that special care be taken in keeping the experiment notebook. This is an important source of evidence for determining how closely the nonexperimental conditions have been kept uniform. Furthermore, in future experimental designs information of this kind will be fundamental in obtaining the next successive higher order of approximation.

Selection and measurement of the criterion. Clearly, no aspect of the experiment is more fundamental than that of setting up and obtaining satisfactory measures of the criterion or criteria against which to evaluate presumed differential effects of the contrasted treatments. We cannot compare alternative treatments until we have agreed upon measures of postulated outcomes acceptable as an index for determining their relative efficacy. Much of the content of this book must be brought to bear on the technical problems encountered here.[1] We can do little more here than to specify them in relation to the requirements of experimental design.

The ultimate criteria of an educational experiment are to be sought in the aims of education themselves which can be determined perhaps only on rational grounds. Such aims are usually stated in broad general terms not amenable to quantitative evaluation. These criteria may be considered ultimate in that it is not possible or practical to go beyond them to seek other goals as educational criteria. Therefore, it becomes necessary for the research worker to set up more direct and immediate criteria if experimentation is to proceed. But it is a part of the task to establish the relation between the ultimate and direct. This may at times be

[1] See especially Chapter IV.

done on the basis of empirical evidence, but more often in the present stage of our knowledge by rational analysis.

Reference has been made earlier in this volume to the necessity for clear statements in operational terms of the objectives in the definition of the problem of investigation. These objectives are the direct criteria in the experiment. Rigorous statements of the objectives in terms of the changes in specific behavior postulated for students as the outcomes of learning experiences in the course of an experiment provide the bases for the construction of the measuring instruments. Rarely are commercial tests available that will suffice for the measurement program. Construction of the necessary instruments is, therefore, ordinarily a fundamental problem for the experimenter. The number of instruments needed is determined by the number of objectives to be measured. The construction of measuring instruments for each of the measurable objectives provides essential information on the effects of the educational factor or factors under inquiry. Thus comparisons can be made of the relative efficacy of the experimental factors with respect to each of the postulated outcomes. The type of measuring instrument to be used will depend upon the nature of the outcome to be estimated. Paper and pencil tests, performance tests, various kinds of anecdotal records, motion pictures, and other types of media may be used depending upon the nature of the functions being tested. Whatever the type of data-gathering devices used, they must meet certain standards as outlined in Chapter IV.

Bias is a matter of great concern in experimentation. If the measurements used are not equally appropriate for the groups under comparison, the true differences between the groups will be contaminated with the systematical effects of bias. Hence no clear cut differentiation from the effects of differential treatment becomes possible. Other sources of bias have been previously pointed out, such as bias in the groups, in the nonexperimental conditions.

Bias is a factor which may result in lowering the validity or reliability of the measuring instrument or of both validity and reliability. Inadequate control of testing conditions, or unequal control, could result in larger variable errors of measurement and of validity in one group than in another, as well as in systematic errors much less readily accounted for. Another important source of bias is the case where the final test is more valid with respect to instruction in one group than with respect to instruction in another group. Of course, if evaluation is made of all possible objectives of instruction in the various groups, such bias does not occur; but in the typical case it occurs quite frequently. For example, consider

the use of the typical standardized test used as the criterion test in an experimental comparison of lecture-demonstration versus laboratory instruction in chemistry.

It should be apparent that the critical experimenter must regard the construction and validation of his measurement instruments as a fundamental problem of the total process of experimentation. He will therefore engage in most of the following technical operations:

1) Construction of the experimental test items for each of the outcomes to be measured.

2) Try out of the experimental items upon a small group or groups representative of the population to which findings of the experiment are to be applied.

3) Analysis of the test data from the sample group. This analysis would include a study of the nature of the distribution of the scores, the mean and standard deviation, estimated reliability, the characteristics of the individual test items, such as their difficulty; correlation with total score as based on each objective (or the grand total), effectiveness of different distractors, and effectiveness of directions and other testing conditions.

4) Preparation of the revised form or forms of the test using the findings from the analysis in (3).

Analysis and interpretation of the experimental observations. Two majoı statistical problems confront the experimenter after he has designed and carried out his experiment. The first problem is that of estimation which consists in applying the method of statistical reduction of the experimental observations which will give the best unbiased estimate of the effect which he presumes to exist. The second is that of applying the appropriate test of significance to determine if a true difference may be inferred from the treatment differences found. The properties of a self-contained experiment have been pointed out as the necessary conditions for the validity of the statistical analysis. These are determined at the time the experiment is designed. The function of randomization has already been discussed.

In experimental work our ultimate interest is in comparing the effects of various methods of treatment. To do so we use certain characteristics of our sample, usually the means, to answer the question whether the difference between the observed values of these characteristics may be attributed to random fluctuations or to the factor(s) under investigation. The null hypothesis states that there is no difference between the effects of the treatments. If

this be true it means that our samples come from the same population and the differences are not statistically significant. A test of significance must be calculated for the difference between the means or other characteristics in which the experimenter may be interested. The most common tests are the *F*-test of the null hypothesis, which assumes that a group of means all have the same true value, and the *t*-test of the null hypothesis, which assumes that a treatment difference between two means is zero or has some other assigned value. Tests of significance should be powerful tests; that is, they should detect the presence of real treatment effects as often as possible. Uniformly, most powerful tests are those that control to the best extent possible what statisticians refer to as the second type of error of judgment, that is, decisions which accept an hypothesis when it is false, or when some alternative hypothesis is true. For example, an experimenter may accept the null hypothesis that there is no difference between the means of the compared methods when in fact there is a difference. Under certain conditions the *F*-test and *t*-test are most sensitive. Under these conditions the experimental errors should be independent, have a common variance, and should be normally distributed. In addition, it is assumed that the assigned and unassignable effects of variation are additive.

It is arbitrary as to how exacting an experimenter may be with respect to the smallness of the probability he would require in order to regard his experimental results as significant. The levels of significance usually set forth are the 5 per cent, the 1 per cent, and 0.1 per cent levels. Whatever the level decided upon, the decision should be made when the experiment is planned. These levels should be considered in relation to the two types of error with respect to the acceptance or rejection of an hypothesis.[1] One of these was referred to above. Sometimes it is equally or more important that we do not reject a hypothesis that has any considerable probability of being true (for example, the null hypothesis referred to above, where there is considerable probability that there is no real difference in the means of the compared methods). Just what balance should be struck will depend upon circumstance. It is the function of statistical tests to show how the errors may be estimated and minimized.

The first problem involved in the analysis of the experimental

[1] See Palmer O. Johnson, *Statistical Methods in Research* (New York: Prentice-Hall, pp. 75–80, 276–326); Max D. Engelhart, "Suggestions with respect to experimentation under school conditions," *Journal of Experimental Education*, 1946, 14:225–244.

data is their reduction to a form which will summarize the information they possess relative to the treatment differences presumed to exist. The choice of the particular summary statistic, which will yield the most relevant information, depends upon the nature of the distribution of the observational data. In random samples from a normal distribution the mean and standard deviation are sufficient statistics. That is, they use all the information available in the sample for estimating the mean and standard deviation of the hypothetical normal curve. Also they may be said to have 100 per cent efficiency. When the test of significance has indicated that a real effect exists, that is, an effect not due to experimental errors, the experimenter may then wish to obtain the estimates with maximum precision. To obtain information about the accuracy of his estimates, fiducial or confidence limits are set up. The fiducial limits represent an interval within which (with a specified fiducial probability) a statement is made that the population value, or parameter, lies. Similarly, there may be determined a confidence interval with a specified confidence coefficient which leads to a statement that the unknown parameter, or, population parameter, will lie within specified limits.

ANALYSIS AND INTERPRETATION OF AN EDUCATIONAL EXPERIMENT

The data from an experiment [1] will be analyzed for illustrative purpose. The experiment was conducted in high school to determine the relative efficacy of two procedures in teaching ninth-grade algebra. Two representative groups of students (that is, representative of the population of this school) comprised the experimental subjects. One group (the control) followed the conventional plan of organization, i.e., they had one fifty-minute class period each school day for two semesters. The other group (the experimental) followed a plan of concentration: they had two fifty-minute class periods each school day for one semester. The nonexperimental conditions were kept uniform: the same teacher taught both groups, using the same method in each group. The students in each pair were randomly assigned to the alternative treatments. The method of obtaining equivalent groups by pairing students on the basis of two basic characters, intelligence quotients (I.Q.) and scores on a prognostic test (Lee Prognostic Test),

[1] For detailed description of the experiment see James L. Hayes, "A comparison of the results of teaching algebra under two conditions: one class period for two semesters and two class periods for one semester," Master of Arts Thesis, University of Minnesota, June, 1938. Only such findings as are necessary for the purpose of the illustration are presented here.

was followed. The criterion, for the part of the experiment illustrated here, was the score on an algebra test (Co-operative Algebra Test). There were 34 pairs of students. The basic data for pairing and the scores of the criterion are presented in Table 25.

It will be noted that on the basis of the summary statistics (the means and standard deviations) the experimental and control groups corresponded closely on the basic characters of matching.

Our interest here is in testing first the statistical hypothesis (the null hypothesis in this case), and secondly in obtaining an estimate of the true difference in the population if the null hypothesis is rejected.

The null hypothesis states that the two groups of measurements (viz., the scores of the experimental and those of the control group) are samples drawn from the same normal population. On the basis of this hypothesis we compare the difference between the means in achievement of the experimental and control groups, with the differences to be expected between these means, in consideration of the observed differences between the achievement scores of individuals of the same group.

The method of pairing, which fixes the details of the arithmetical procedure so that clear-cut interpretation of the experimental observations may be made, involves a number of assumptions. The method of pairing is assumed to have equalized any differences in intelligence, and in other factors indicative of achievement, in which the several pairs of individual students may differ. These differences, which have been removed from the experimental comparisons and have therefore not contributed to the real errors of the experiment, must be eliminated in a like manner from the estimate of error. It is upon the basis of the estimate of error that a decision must be made as to what differences between the means are consistent with the null hypothesis and what differences are incompatible with the hypothesis. The primary concern is not with the differences in achievement among the students in the same group but with the differences in achievement between students who are in the same pair, and with dissimilarities among the differences found in the different pairs. The first step in the statistical analysis, then, consists in subtracting the score on the criterion of each student in the experimental group from the score of each student in the control group belonging to the same pair. These differences are shown in Column (8), Table 25.

The null hypothesis in terms of these differences states that they are distributed in a normal manner about a mean value at zero. We wish to test the second aspect of the hypothesis, viz., that our sample of 34 observed differences may be regarded as a random

TABLE 25. Comparative Measures on Basic Characters of Matching and Scaled Scores on Co-operative Algebra Test of the Experimental and Control Groups

Number	*Intelligence Quotients*		*Lee Prognostic Scores*		*Scaled Scores on Algebra Test*		*(X)* $E - C$
	Control	Exper.	Control	Exper.	Control	Experimental	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	143	135	119.2	116.7	66	72	6
2	140	135	105.3	117.2	57	64	7
3	134	132	131.8	133.2	64	85	21
4	132	132	129.5	123.3	74	67	−7
5	131	137	77.5	86.0	63	73	10
6	130	128	126.3	119.7	68	70	2
7	122	121	95.9	98.0	65	67	2
8	122	123	113.3	113.9	54	67	13
9	120	120	107.1	110.8	58	55	−3
10	118	118	78.4	77.6	46	48	2
11	117	114	59.9	57.9	45	55	10
12	117	118	79.5	82.5	57	52	−5
13	116	117	88.5	90.2	60	70	10
14	116	117	85.8	87.9	52	64	12
15	116	117	70.9	67.0	60	52	−8
16	114	111	100.1	96.8	52	54	2
17	114	114	115.5	110.3	64	71	7
18	114	114	88.4	89.0	53	60	7
19	112	109	53.3	54.3	44	50	6
20	108	107	85.5	85.2	46	42	−4
21	109	108	73.6	71.1	48	52	4
22	108	106	95.9	94.0	40	54	14
23	106	107	87.4	89.5	56	53	−3
24	106	106	106.0	105.8	49	64	15
25	105	105	70.0	65.0	59	50	−9
26	105	106	116.4	117.3	58	70	12
27	103	102	60.0	53.1	51	55	4
28	102	100	82.0	76.0	57	51	−6
29	100	111	105.4	104.1	58	64	6
30	100	100	104.9	112.5	53	55	2
31	100	100	54.4	56.6	55	51	−4
32	100	96	93.9	98.2	41	59	18
33	97	92	79.4	79.2	52	58	6
34	96	86	67.5	64.6	52	52	0
Mean	113.9	113.06	91.4	91.3	55.21	59.59	4.38
S.D.	12.36	12.58	21.54	22.07	7.73	9.198	1.31

sample from the population with a true mean difference of zero. If we wished to test out the first part of the hypothesis, "in a normal manner," a specific test of normality would be essential. This second test does not concern us here.

To test the null hypothesis that our random sample of differences comes from a population distributed about a mean of zero, we need to calculate a value of a criterion known as *t*, which in this problem may be defined as the mean of the individual pair differences divided by the standard error of the mean as estimated from the number of degrees of freedom. The number of degrees of freedom is the number of independent relations existing among the 34 differences, which is in this case 34 – 1 or 33.

We need then to calculate the mean, $\bar{x}$, of the individual pair differences, which is their sum, 149, divided by 34:

$$\bar{x} = \frac{149}{34} = 4.38$$

$$\begin{aligned} \text{also } \bar{x} &= \bar{x}_{exp.} - \bar{x}_{con.} \\ &= 59.59 - 55.21 \\ &= 4.38 \end{aligned}$$

We also need to compute the standard error of the mean, $s_{\bar{x}}$. For this computation, we find first the variance of the individual differences and then divide this variance by the number of observations. Then we take the square root of this ratio, and we have the standard error of the mean.

The variance, s_x^2, of the individual paired differences for this problem is obtained by first calculating the sum of squares of the several differences, deducting from this sum the quotient of the square of the sum of the individual pair differences by their number, and dividing the difference by the number of degrees of freedom. Thus

$$\begin{aligned} s_x^2 &= \frac{\Sigma(X - \bar{X})^2}{N - 1} \\ &= \frac{\Sigma(X^2) - \dfrac{(\Sigma X)^2}{N}}{N - 1} \\ &= \frac{2591 - \dfrac{(149)^2}{34}}{33} \\ &= \frac{1938.03}{33} \\ &= 58.728 \end{aligned}$$

The variance of the mean $s^2_{\bar{x}}$ is obtained by dividing the variance of the individual paired differences, s^2_{x}, by the number of pairs, N. Thus

$$s^2_{\bar{x}} = \frac{s^2_x}{N}$$
$$= \frac{58.728}{34}$$
$$= 1.7273$$

The standard error of the mean, $s_{\bar{x}} = \sqrt{\frac{s^2_x}{N}}$

$$s_{\bar{x}} = \sqrt{1.7273}$$
$$= 1.31$$

The standard error of the mean can be obtained in one calculation as

$$s_{\bar{x}} = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N(N-1)}}$$
$$= \sqrt{\frac{58.728}{(34)(33)}}$$
$$= \sqrt{1.7273}$$
$$= 1.31$$

The value of t_o is, then

$$t_o = \frac{\bar{x}}{s_{\bar{x}}}$$
$$= \frac{4.38}{1.31}$$
$$= 3.34$$

The purpose of these calculations has been to obtain from the observational data a quantity measuring the mean difference in achievement between the control and experimental paired individuals in terms of the observed discrepancies among these differences. The distribution of the quantity in repeated sampling must be known under the assumption that the null hypothesis is true. The mathematical distribution of the quantity, t, known as the "t" distribution, is dependent only on the number of degrees of freedom available for calculating the estimate of error.

In our problem, the number of degrees of freedom is 33. We enter the "t" table then, to answer the question: What is the probability that random sampling could give a value of t deviating from

the true value, or $t = 0$, by an amount equal to or greater than $\pm t_o$ as found in our problem. We enter this table with 33 degrees of freedom and the probability value corresponding to a t of ± 3.34, as found in our problem is $< .01$.

Since the probability, then, is so small, less than one in a hundred trials, that random sampling could give a value of $t_0 \geqq \pm 3.34$, we reject the null hypothesis that our random sample of individual pair difference came from a population of differences distributed about a mean of zero.

In other words, the hypothesis that there is no difference between the two organizations of teaching high-school algebra has been refuted by the facts of observation.

Having established from the test of significance that the true difference of the mean individual pair differences is not zero, the experimenter may now solve the problem of estimation by setting up the fiducial limits or confidence interval for the true mean difference.

The fiducial limits are obtained in the following manner:

We may write, where μ is the true difference in the population

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \qquad \text{(1)}$$

$$t = \frac{4.38 - \mu}{1.31} \qquad \text{(2)}$$

Then, for a fiducial probability of 99 per cent, for instance, we determine from the table of the t distribution that the value of t is 2.733 for $\mu = N - 1 = 33$.

Inserting the values of $\pm t_{.01} = 2.733$ in equation (2) we can solve for μ. Thus

$$\begin{aligned} \pm 2.733 &= \frac{4.38 - \mu}{1.314} \\ \mu &= 4.38 \pm 3.59 \\ &= 0.79, \text{ or } 7.97 \end{aligned}$$

We may then say that the fiducial probability is 99 per cent that the true mean difference μ will be within the fiducial limits 0.79 and 7.97.[1]

The confidence interval is set up as follows. For a confidence coefficient of 99 per cent $t_{.01} = \pm 2.733$.

The upper and lower limits of the confidence interval are

[1] Palmer O. Johnson, *Statistical Methods in Research* (New York, Prentice-Hall), p. 115.

$$\bar{X} + t\epsilon\sqrt{s^2/N} \text{ and } \bar{X} - t\epsilon\sqrt{s^2/N}$$

$$\bar{X} + t\epsilon\sqrt{\frac{s^2}{N}} = 4.38 + 2.733\sqrt{\frac{58.728}{34}}$$

$$= 7.97$$

$$\bar{X} - t\epsilon\sqrt{\frac{s^2}{N}} = 4.38 - 2.733\sqrt{\frac{58.728}{34}}$$

$$= 0.79$$

The confidence interval is (0.79, 7.97).

We can make the statement that the interval extending from 0.79 to 7.97 will cover the true mean individual pair differences and we know that we shall be right in 99 per cent of such cases.[1]

MODERN EXPERIMENTAL DESIGNS

The experimental model. The principles of experimentation are capable of wide application. It is not within the scope of this volume to present in detail modern developments in experimental design. We shall, however, conclude our discussion by summarizing the main characteristics and the significance of this development for educational research.

When one desires to treat some real problem mathematically, whether in the physical, the biological, or the social sciences, one must necessarily begin by simplifying the problem, having recourse to some kind of a model representing those features regarded as most important for the problem in question.

The classical model for the ideal experiment was built upon the concept of varying only one factor at a time, the other conditions being kept as uniform and constant as laboratory conditions would permit. The tradition of this model of experimental design has been handed down from its famous origin in physics as Galileo's investigation of the basic laws of falling bodies. Galileo's model was accepted as appropriate by workers in the biological field until R. A. Fisher and others designed the modern model of experimentation. Instead of isolating single factors for investigation, the basic intent of this model was to give full play to the factors which arise in practice in order to study what takes place in "natural" situations. Accordingly, modern experiments are frequently complex

[1] For the statistical treatment see Palmer O. Johnson, *op. cit.* Ordinarily either the fiducial limits or the confidence interval is computed. Since the student will encounter both in the literature, both are illustrated here. They do not always come to the same thing.

in that a number of factors are introduced simultaneously into the same inquiry. The effects of each factor are determined as well as the effects of the interactions of the several combinations of factors.

The principles of experimentation formulated by Fisher were well established by 1926. They covered such essentials of efficient experimental design as replication, randomization, and control of variability. Appropriate and efficient statistical tools for the analysis and interpretation of experimental results had already been provided by the technique of analysis of variance first reported by Fisher in 1923. These were a necessary condition for modern experimentation. The analysis of variance technique provides the appropriate method of estimating the experimental error and of carrying out the exact tests of significance. We shall not discuss the technical treatment of these processes. The reader is referred to the references of this chapter. We shall, however, discuss briefly the role played by the analysis of variance.

In the original model, each experimental observation is represented as the sum of a number of components usually four assigned, respectively, to the general mean, the effect of the treatments under comparison, certain environmental effects which the design of the experiment makes it possible to isolate, and the residual effect representing the measure of all other sources of variation that might affect the observations. This component is generally referred to as "experimental error."

The role of analysis of variance and covariance. In like manner the analysis of variance partitions the total sum of squares of the deviations of the observations from the grand mean into four different sums of squares, one assigned to the general mean (the general average about which it is presumed that the observational values fluctuate), a second attributable to the difference between the estimated effects of the treatments, a third to the environmental effects which the experimental design makes capable of measurement, and the fourth, which is the residual or error sum of squares. In practice the procedure is usually to calculate the original sum of squares and the first three components. The error sum of squares is then obtained by subtraction.

The technique thus makes possible the analysis of the experimental results such that the respective components of variation, which the design was intended to identify, may be isolated and the effects measured. Of greatest importance is that the effects of the several groups of observations, which have been measured, are eliminated from the estimates of treatment effects. If this separation

were not possible, these differences would inflate the value of the experimental error with the result that less accurate estimates of the treatment effects would be secured.

The role of the analysis is not restricted to that of providing a short-cut means of obtaining the error sum of squares. The sum of squares attributable to treatments is the quantity required for the *F*-test of the null hypothesis that no differential effects exist between the different treatment effects. The sum of squares due to environmental effects can be used to estimate the increased precision of the experiment, which has been achieved by eliminating the environmental effects from the estimates of the treatment-effect means.

A complete analysis of variance is shown in Table 27. This table presents both the results of the analysis in a compact simple form as indicated by the break-up of the total sum of squares and the logical structure or design of the experiment as shown by the division of the total number of "degrees of freedom." There is associated with each component a certain number of "degrees of freedom," which is a technical term for the number of independent parameters needed to specify that particular component in the experimental model. In the case of treatments (e.g., "method" in Table 27), the number of degrees of freedom is one less than the number of treatments; similarly, for the component, the controlled environmental factors (e.g., the "schools"). The number for the total sum of squares is the total number of observations less one, which represents the contribution of the general mean.

There are two principal uses of the degrees of freedom. The first is that by subtraction they give the number of degrees of freedom for error. This number is the divisor required for the error sum of squares in order to estimate the population variance, σ^2. Secondly, the degrees of freedom for treatment are required in an F-test of the null hypothesis that all treatment effects are the same. Likewise, the F-test may be used to test the significance of the effects of the other components, if desired.

Reference has been made to the technique of the analysis of covariance. Further information on this technique may be found in the references.[1] We may say here, however, that it is another method of achieving reduction in the experimental error and thereby increasing the precision of the experiment. It utilizes an-

[1] Max D. Engelhart, "Suggestions with respect to experimentation under school conditions, *Journal of Experimental Education,* 1946, 14:225–244; Palmer O. Johnson, *Statistical Methods in Research* (New York: Prentice-Hall, pp. 75–80, 276–326); E. F. Lindquist, *Statistical Analysis in Educational Research* (Boston: Houghton Mifflin Company, 1940, pp. 76–86).

cillary information in such a way that certain types of environmental effects are eliminated from the estimates of treatment effects and thereby rendering these estimates more accurate. For instance, if we measure the growth in achievement of pupils under contrasting treatments, the potential growth rates are likely to be different at the beginning of the experiment because of inequalities in intelligence of the experimental subjects. Therefore, the two or more groups of pupils under contrast are likely to differ in achievement from this cause alone. However, we may obtain measurements on an intelligence test of all the experimental subjects. These measures provide ancillary information that may be taken into account in improving the accuracy of the estimates of achievement growth rates under the contrasting treatments.

In general, then, we note that in addition to the experimental variables available for analysis, we may have another value related to the individual pupils but unaffected by the treatment given to them. This is the simplest case. There may be, in general, a number of such values. For example, in addition to the intelligence test score, we may have the initial score on the criterion, or other measures of aptitude related to it. In this way, sources of variation not amenable to control by the experimental design often can be measured by taking additional observations.

Accordingly, we may expect to find the precision of our comparison increased provided (a) the ancillary measure is reasonably highly correlated with the experimental variable after allowing for the fact that the latter may differ because of being differentially treated and (b) we are justified in correcting the experimental variable for differences in the matching variable, i.e., in adjusting the values on the criterion for the different treatments to correspond to equal values of the ancillary variables (a variable unaffected by those treatments).

The technique of analysis of covariance shows how to use the supplementary data by eliminating effects of variation in the ancillary variable (*s*). It is essentially a technique of regression analysis. With one adjusting variate we have simple regression—with more than one we enter the realm of partial and multiple regression.

Laboratory versus "natural" conditions. The change from the classical notion of experimentation, with its emphasis upon a single experimental variable, to the modern ideas of combining several lines of inquiry in a single large-scale experiment, or, the shift from the univariate to the multivariate case has altered our point of view with respect to the role of the laboratory in experimentation. No longer can the laboratory be regarded as a necessary con-

dition for controlled experiment. That is, it can no longer be regarded as obvious that laboratory techniques are *sufficient* for securing answers to critical scientific questions, not excepting the physical sciences, nor that the laboratory is *necessary* for collecting a set of valid and reliable data on a given educational question.

Many persons maintain that reliable data must be obtained under the conditions of actual service or use. They insist that information must be obtained under the most general conditions in which the theory under test is expected to apply. On the other side, some technologists believe that service conditions are in many instances as highly specialized as those of the laboratory and that the only distinction lies in the fact that all the conditions in the service test are not known.[1] However, modern experimental designs, including the correspondingly appropriate methods of statistical analysis, have shown how to resolve this conflict between the laboratory tests and the "natural" conditions. The modern research worker through use of these modern tools is now able to obtain the *stochastic* relation between variables under the conditions observed in practice.

We thus note the current transition in science of a 19th-century concept: the procedure of abstraction whereby the scientist may view his laboratory as an isolated system shutting out the outside world with its uncontrollable variables, the uncertainties which would not submit themselves to quantification. This device of abstraction is particularly difficult in the case of the living organism. It is difficult because the living organism is essentially adaptive and the problem for study is compounded of the organism and its situation. To isolate the organism from its setting by transporting it to the laboratory may rupture the problem.[2]

The impact of these modern principles upon experimentation is now beginning to be observed in many fields. Egon Brunswick, for example, has singled out the field of perception and by a sequence of four experiments, which involve threshold, illusion (Gestalt dynamics), constancy of apparent size, and social perception of intelligence and personality traits under conditions of restricted contact, traces the progress in methodology in experimental psychology over a period of a century.[3] In particular, he uses the series to

[1] West C. Churchman, *Theory of Experimental Inference* (New York: The Macmillan Co., 1948), 292 p.

[2] Ewen D. Cameron, "The current transition in the conception of science," *Science, 107:* 553–558.

[3] Egon Brunswik, "Systematic and representative design of psychological experiments with results in physical and social perception," *Proceedings of the Berkeley Symposium* (University of California Press, 1949), pp. 143–207.

illustrate the transition in progress from a "classical" style of laboratory research to the field type of experiment which has been influenced by modern statistical developments.

Modern ideas of experimental design are being recognized in educational experimentation. At least there has been sufficient trial to indicate promise for modern methods of experimentation in this field. Particularly promising both in educational and psychological experimentation is the possibility of extending the generalizability of experimental evidence not only by obtaining a representative sampling of experimental subjects but also by obtaining representative objects and/or representative stimulus situations in the experimental design.

Of special promise is the development of factorial designs [1] in which the effects of a number of different factors are investigated simultaneously. The treatments are comprised of all the possible combinations of the different factors. Of particular importance is the investigation of the interactions among the effects of several factors. The factorial experiment is especially appropriate when recommendations are made relative to a wide range of conditions. In this type of experiment the principal factors can be evaluated under a variety of conditions similar to those encountered in the population to which the recommendations are to be applied.

Modern experimental research has led to wide-spread co-operative effort in the field of education. By co-ordinating research in several different schools, one not only obtains the value of each of the individual researches, but also the additional knowledge accruing from the combined results. The advantage of modern designs resides in their capacity to recognize the genuineness of discrepant findings obtained at different times and places. Thus one notes the limitation of generalizations which might otherwise be accepted uncritically.

A modern experiment illustrated. We shall conclude this discussion of modern experimental designs by presenting as an illustration the application made by Burt and Lewis [2] in their study of

[1] William C. Cochran and Gertrude M. Cox, *Experimental Designs* (New York: John Wiley and Sons, 1950).

Loretta E. Heidgerken, "An experimental study to measure the contribution of motion pictures and slidefilms to learning certain units in the course introduction to nursing arts," *Journal of Experimental Education*, 1948, 17:261–281.

Palmer O. Johnson, *Statistical Methods in Research* (New York: Prentice-Hall), pp. 75–80, 276–326.

E. F. Lindquist, *Statistical Analysis in Educational Research* (Boston: Houghton Mifflin Company, 1940), pp. 76–86.

[2] Sir Cyril L. Burt and Bernard Lewis, "Teaching backward readers," *British Journal of Educational Psychology*, 1946, 16:116–132.

the problem of remedial instruction in reading. Only one part of the comprehensive study[1] is reported. The part with which we shall be concerned had as its special point of interest the determination of the effects of previous teaching methods upon the success of the method subsequently used in remedial teaching. The type of experimental design used is that known as the *randomized block design*. In this design the experimental area consists of a number of blocks and each block is subdivided into a number of plots. In each block a treatment is assigned at random to one and only one plot so that each treatment appears in each block only once.

There were 48 experimental subjects: boys aged 10 years, 8 months to 11 years, 3 months, with I.Q.'s between 79 and 83 and reading quotients between 61 and 73. The plan of the experiment is shown in Table 26. It is noted that there are two criteria of classification according to (1) the method used in remedial work; i.e., the alphabetic, kinaesthetic, phonic, and visual (e.g., the treatments); and (2) the teaching method to which the child had previously been taught in his own school (labeled "schools" in Table 26, corresponds to "blocks").

There were accordingly 16 subclasses and there were three pupils in each subclass. Since interaction between method and school was a principal concern in the investigation three determinations of the variate were provided in each subclass. The treatments were assigned at random, one treatment to each block.

The instructional materials were the same for all the methods, that is, the reading-matter, words, sentences, stories, etc. were the same, and with each method the commoner devices for arousing interest were adopted (e.g., free use of pictures, games, puzzles, and stories).

The values of the criterion in the body of Table 26 represent improvement ratios, i.e., the ratio obtained on dividing the pupil's final achievement quotient by his initial achievement, or: $\frac{\text{R.A.}(2)}{\text{R.A.}(1)} \times \frac{\text{M.A.}(1)}{\text{M.A.}(2)}$.

The analysis of variance is given in Table 27. The tests of significance and the calculations required for the total and the several component sums of squares are given following the table.

The tests of the various hypotheses, made by the *F*-test lead to the following conclusions:

[1] The reader is referred to the original article for all the details. A preliminary experiment dealt with the evaluation of the effects of method, age, and sex of the instructor. Only as much report is made here as seems necessary to illustrate the principle underlying the experimental design used.

TABLE 26. Improvement Ratios of Pupils in Reading: Second Experiment

	Schools					
Methods	*A*	*B*	*C*	*D*	*Total*	*Average*
A. Alphabetic	98.7	114.2	102.8	103.2		
	109.4	101.6	96.1	111.5		
	100.2	113.0	110.2	106.4		
Total	308.3	328.8	309.1	321.1	1267.3	105.6
Average	102.8	109.6	103.0	107.0		
B. Kinesthetic	118.6	102.5	113.5	103.9		
	106.0	106.7	117.4	119.1		
	116.1	110.4	116.8	116.2		
Total	340.7	319.6	347.7	339.2	1347.2	112.3
Average	113.6	106.5	115.9	113.1		
C. Phonic	107.5	106.4	111.2	101.6		
	112.6	98.4	100.8	105.5		
	105.3	93.6	109.6	94.7		
Total	325.4	298.4	321.6	301.8	1247.2	103.9
Average	108.5	99.5	107.2	100.6		
D. Visual	128.1	113.4	119.8	101.2		
	119.0	119.2	106.6	108.0		
	126.5	111.3	107.9	107.6		
Total	373.6	343.9	334.3	316.8	1368.6	114.0
Average	124.5	114.6	111.4	105.6		
Total	1348.0	1290.7	1312.7	1278.9	5230.3	
Average	112.3	107.6	109.4	106.6		108.9

There was a significant difference among the means of the methods leading to the rejection of the hypothesis that all the methods of instruction produced the same effect. $F = 9.31$ is greater than 4.46, the F required for significance at the 1 per cent level.

There was no significant difference among the means of the schools, that is, the hypothesis that all the school effects are identical was accepted. $F = 2.44$ is less than 2.90, the F required for significance at the 5 per cent level.

There was a significant interaction effect between schools and method (at the 5 per cent level). $F = 2.70$ is greater than 2.19, the F required for a significance at the 5 per cent level. This may be interpreted that with many of the pupils the improvement made in reading with any of the given remedial treatments depended on

TABLE 27. Analysis of Variance of the Improvement Ratios Recorded in Table 26

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Squares*
Methods	3	880.2	293.4
Schools	3	230.6	76.9
Interaction (Schools x Methods)	9	763.7	84.9
Error (Within subclasses)	32	1008.7	31.5
Total	47	2883.2	

F-ratio	*D.F.*	$F_{.05}$	$F_{.01}$
Methods: 293.4/31.5 = 9.31	3,32	2.90	4.46
Schools: 76.9/31.5 = 2.44	3,32	2.90	4.46
Interaction: 84.9/31.5 = 2.70	9,32	2.19	3.01

Computation of the sums of squares listed in Table 27.

1) Correction: $\frac{(5230.3)^2}{48} = 569917.46 = C$

2) Total: $(98.7)^2 + (109.4)^2 \cdots\cdots\cdots + (107.6)^2 - C = 2883.2$

3) Subclasses: $\frac{(308.3)^2 + (340.7)^2 \cdots + (316.8)^2}{3} - C = 1874.5$

4) Within subclasses: (2) − (3) $= 1008.7$
(error)

Can verify the last result by adding the 12 sums of squares obtained from applying the usual computational method to each of the subclasses. In the first subclass, for example, the sum of squares is:

$$(98.7)^2 + (109.4)^2 + (100.2)^2 - \frac{(308.3)^2}{3} = 67.1$$

The sum of squares for subclasses is now divided into three parts:

5) Methods: $\frac{(1267.3)^2 + \cdots + (1368.6)^2}{12} - C = 880.2$

6) Schools: $\frac{(1348.0)^2 + \cdots + (1278.9)^2}{12} - C = 230.6$

7) Subclass discrepancy
(treatment interaction): (3) − (5) + (6) = 763.7

The mean squares in Table 27 are obtained by dividing the sum of squares in each row by the corresponding number of degrees of freedom.

the method by which the child had originally been taught. On the whole, a change in method of remedial instruction appeared more effective than renewed efforts with an old procedure with which the child was already familiar.

This experiment is of special interest in that it makes possible conclusions regarding the interacting effect. This is one of the important contributions of modern designs. This effect could not have been detected by the traditional single factor type of experiment.

The critical reader may have noted that the mean square due to error was used in all the tests of hypotheses. This was the proper basis since the schools were not chosen at random, although the pupils were chosen at random from the schools included. The fact that the schools were not chosen at random restricts the conclusions to the particular schools included.

If the schools had been chosen at random from the population of schools (all London schools, for example) then the appropriate mean square to use in the F-tests for methods and schools would have been that due to interaction. The broader generalization for all London schools would have required that the schools included in the experiment had been chosen at random from the London schools.

SUMMARY

The experimental method is essentially a means of gaining new knowledge by the collection of fresh observations under controlled conditions. This method thus adds enormously to the scope of scientific inquiry in that the investigator is no longer limited to observing nature's experiments or phenomena that occur in the natural, social, economic, or educational environment.

If we are to secure with the certainty possible in science, estimates (in the causal sense) of the effect of any given factor or combination of factors, experiments must be undertaken. In experimentation the purpose is to set up a relatively simple system of causes and effects so that observations on the system constitute data capable of analysis by the ordinary principles of the logic of scientific method and by the appropriate statistical procedures.

We have considered in this chapter the comparative experiment, which consists of the application and comparison of differential treatments. The process of adding to scientific knowledge by experiment consists of the following sequence of events: (1) critical examination of theories on the basis of available evidence; (2) the formulation of hypotheses that are testable or appropriate for testing by experimentation; and (3) the carrying out or execution of experiments.

The ideas underlying modern design of experiments differ

sharply from traditional ones. The traditional experiment was based upon the formula: Establish controlled conditions such that all factors except one can be held constant and then study the effects of this single factor. Modern statistical science has repudiated the assumption that in order to apply the experimental method it is necessary to keep all conditions constant with the exception of the experimental factor under study. It has been shown through the application of the experimental principles of replication, randomness, and control of variation that an experimental problem can be investigated when several variables are undergoing change.

CHAPTER IX

The Problem of Prediction

Simple regression. The practical man depends upon empirical relationships to aid him in estimating or predicting one characteristic from a knowledge of another characteristic, or the value of one quantity from that of another. Sometimes the problem may be that of assessing the value of some quantity which may be difficult or impossible to observe directly in a given instance. Thus the physiologist might be interested in the estimation of brain weight or heart weight from body weight. At other times the relationship between two or more quantities may aid the practical man in setting up a procedure which has good prospects of achieving a needed outcome. Thus a teacher may wish to estimate achievement or to predict future achievement of his pupils from a knowledge of their I.Q.'s, with a view to adjusting instruction to individual differences. Underlying all evaluation, prediction or estimation is involved. When a teacher constructs an achievement test in English, he implies that the score obtained by a student is an estimate of his total achievement in this field.

In current scientific practice data are collected for action—action with respect to some problem that prompted their collection. Such action to be based on scientific grounds presupposes a certain knowledge of future events based upon propositions which have been subjected to scientific test for their validation in the past. The predictional value is thus the basis for current and future action. Whether there exist statistical bases for the prediction of unknown events depends upon the state of scientific knowledge.

One means for making scientific prediction is the algebraic equation summarizing an ascertained relationship among measurable

variables. The formulation of such equations involves a careful study of the phenomena under investigation and the use of special statistical techniques that we wish now to discuss. We shall first illustrate the solution of a simple problem in prediction, which is of frequent occurrence in education.

We begin by defining the population to be sampled, which, in our case, is a group of graduate students in educational psychology. The population sampled should coincide with the population about which the information is desired. We wish to predict the achievement of graduate students on a midquarter examination in educational psychology from a knowledge of their scores on an intelligence test, namely, Miller's Analogies. When we predict one characteristic from another characteristic, as in this example, we presume that a change in one variable is accompanied by some corresponding change in the other variable, or that a certain relationship exists between the two variables. We are concerned here with the extent to which two quantities vary together. We wish to reduce this tendency to a rigorous quantitative basis in order to obtain all the information in the data to predict the mid-quarter score from the score on the Analogies Test. The problem involves the development of an appropriate regression equation. Here we have scores on Miller Analogies and scores on the midquarter examination for a representative sample of students. If now we were given the analogies scores only of another sample of students from the same population, how accurate a guess could we make of their midquarter scores on the basis of the available information?

The data are presented in Table 28 for a random sample of 25 students. The students were drawn, one by one, without replacement. At every state, or every draw, all undrawn students had an equal chance of selection. This process was applied in the selection of our sample. A table of random sampling numbers was used. It is assumed that the conclusions reported here are for the population of graduate students to be taught in the same way and with the same materials as those constituting the sample which served to provide the observational results.

The most obvious way of determining whether any relationship exists between the two variables under consideration is that of plotting the observed pairs of values on graph paper. This has been done in Figure 4, where the scores of the 25 students on the midquarter and Miller's Analogies have been plotted with reference to two co-ordinate axes, the Y-axis according to convention allotted to the dependent variable, midquarter score. It is apparent that the points are not scattered at random over the XY-plane, but appear

TABLE 28. Calculation of the Regression Equation for Midquarter Examination Scores (Y) on Miller's Analogies Test Scores (X) for a Random Sample of 25 Graduate Students

X	Y	X'	Y'	X'^2	Y'^2	$X'Y'$
85	82	25	12	625	144	300
47	66	−13	−4	169	16	52
67	89	7	19	49	361	133
75	97	15	27	225	729	405
53	86	−7	16	49	256	−112
56	56	−4	−14	16	196	56
75	95	15	25	225	625	375
54	64	−6	−6	36	36	36
43	69	−17	−1	289	1	17
99	103	39	33	1521	1089	1287
76	70	16	0	256	0	0
65	52	5	−18	25	324	−90
79	81	19	11	361	121	209
69	72	9	2	81	4	18
78	84	18	14	324	196	252
67	78	7	8	49	64	56
75	75	15	5	225	25	75
73	75	13	5	169	25	65
72	67	12	−3	144	9	−36
51	74	−9	4	81	16	−36
79	86	19	16	361	256	304
67	83	7	13	49	169	91
74	82	14	12	196	144	168
96	98	36	28	1296	784	1008
84	85	24	15	576	225	360
1759	1969	259	219	7397	5815	4993

$X' = X - 60 \qquad \bar{X} = 60 + \frac{259}{25} = 70.36$

$Y' = Y - 70 \qquad \bar{Y} = 70 + \frac{219}{25} = 78.76 \qquad Y_E = \bar{Y} + \frac{\Sigma xy}{\Sigma x^2}(X - \bar{X})$

$\Sigma x^2 = 7397 - \frac{(259)^2}{25} = 4713.76 \qquad = 78.76 + \frac{2724.16}{4713.76}(X - 70.36)$

$\Sigma y^2 = 5815 - \frac{(219)^2}{25} = 3896.55 \qquad = 78.76 + .5779X - 40.6622$

$\Sigma xy = 4993 - \frac{(259)(219)}{25} = 2724.16 \qquad Y_E = .5779X + 38.0978$

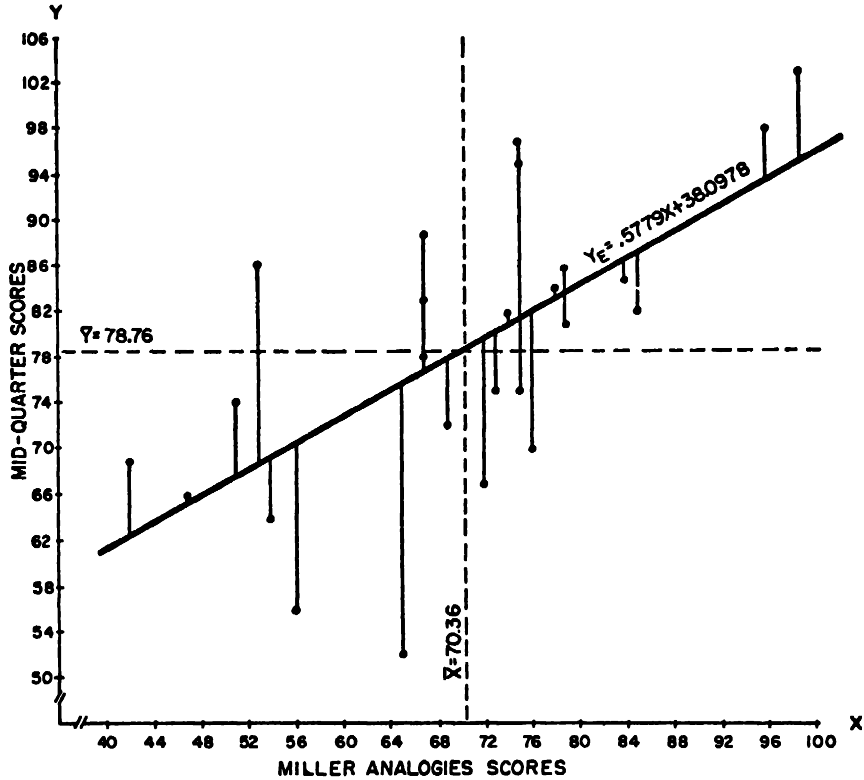


FIGURE 4. Line of best fit in least squares sense, fitted to data in Table 28. (Graph for equation: $Y_E = .5779X + 38.0978$)

to be arranged in a band extending from the lower left-hand to the upper right-hand corner. There is an apparent tendency for students scoring high on the Analogies test to score high on the mid-quarter, for those who score low on the one to score low on the other, and for those who are average on the one to be more or less average on the other.

Although the general trend of the plotted points may suggest the presence of a relationship, the plotted points do not themselves provide a precise expression of that relationship. Some means is needed to express this relationship concisely. Moreover, in observational science data are subject to all kinds of fluctuations. Hence, trends that appear to exist may not upon closer examination be found to be real. Means must be provided for testing the validity of the inference drawn from the observations concerning the existence of a true relation between the quantities concerned. Particularly, in biological and psychological data, in addition to fluctu-

ations arising from sampling and from the inaccuracies of measurement by experimental or observational techniques, there is also the real variability which is an essential part of the phenomenon studied. The complete analysis of the problem of regression is somewhat involved if one is to guard against unwarranted conclusions.

The observation that the plotted points in the diagram seem to be somewhat orderly arranged in a direction from the lower left-hand to the upper right-hand corner suggests that, although the locus is not sharply outlined, a straight line might be drawn among them. The determination of the slope of this line should enable us to quantify the tendency for scores on one measure to change with scores on the other measure. The slope is the ratio of the opposite to the adjacent side of the angle which the line makes with the positive direction, say OX on the X-axis.

We might proceed, as Galton did in his study of regression, by drawing a line (by inspection) among the 25 plotted points and determine its slope by graphic means. This method, however, would not lead to precise results since the slope of the line would depend upon the judgment of the individual making the inspection. To solve this problem objectively we make use of the well-known *principle of least squares,* which is used in various ways in analytical work. Here the principle is to choose that regression line whose slope is such as to minimize the sum of squares of the vertical projection of each point on the line. This line is frequently referred to as the line of "best fit." The projections are shown in the figure. It is assumed here that the regression is linear. This assumption can and should be tested.[1]

Computation of regression equation from raw scores. The method of calculating the slope of the "best" straight regression line in the sense of the method of least squares is carried out in Table 28.

The equation of a straight line in intercept form is:

$$Y_E = a + bx \quad \ldots\ldots (1)$$

In our case this may be written

$$Y_E = \bar{Y} + \frac{\Sigma xy}{\Sigma x^2}(X - \bar{X}) \quad \ldots\ldots (2)$$

The required values are the means $\bar{Y}$ and $\bar{X}$ and slope, b, of the line given by:

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

[1] See, for example, Palmer O. Johnson, *Statistical Methods in Research* (New York: Prentice-Hall, Inc., 1949), p. 240.

where x and y are deviations from the mean, and Σx^2 is the sum of squares of the deviations of the X values from their mean. In calculating the slope from the original measures rather than from deviation scores the formula is written:

$$b = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma X^2 - \frac{(\Sigma X)^2}{N}} \qquad (3)$$

where ΣXY is the sum of products of the raw scores; $\overline{Y}$, the mean of Y, and $\overline{X}$, the mean of X are determined in the usual way. To reduce the amount of labor the original measures have been first reduced by subtracting 60 from the X scores, and 70 from the Y scores.

The regression equation of Y on X was found to be

$$Y_E = .5779X + 38.0978 \qquad (4)$$

where Y_E denotes the estimated value of Y, or, the estimated mid-quarter score.

The equation tells us how changes in Y vary with unit changes in X.

The estimated values (Y_E's) for the 25 students are given in Table 29. We can at any subsequent date use this equation to predict a student's score on the midquarter exam given his score on Miller's Analogies (for example, see Table 30).

Further analysis, however, will be made by the critical worker, who is aware of the hazards involved in accepting findings at their face value.

Obviously, the importance of the prediction equation lies in its subsequent prediction for other groups upon whom only the knowledge of the one characteristic is available. We must know, then, the confidence we may place in its use for such purposes.

We must note first that equation (1) is merely an estimate of the population regression equation which may be written:

$$\bar{Y} = \alpha + \beta (X - \bar{X}) \qquad (5)$$

where α and β represent the population mean of the Y's and the population regression coefficient, respectively. In the practical case under consideration we are testing the hypothesis that $\beta = 0$, i.e., that there is no regression of Y on X in the population sampled. For testing this hypothesis from the sample data, we test the significance of our obtained regression coefficient: $b_{yx} = .5779$, that is, we test whether b_{yx} is significantly different from zero. For the required test we calculate the value of

$$t_o = \frac{b_{yx}}{\sigma_b} \quad \ldots\ldots (6)$$

where σ_b is the standard error of b given by $\frac{\sigma_{y.x}}{\sqrt{\Sigma x^2}}$

$\sigma_{y.x}$ is the standard error of estimate and is obtained from

$$\sigma_{y \cdot x} = \sqrt{\frac{\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}}{N - 2}} \quad \ldots\ldots (7)$$

$$= \sqrt{\frac{3896.55 - \frac{(2724.16)^2}{4713.76}}{25 - 2}}$$

$$= 10.048$$

Recalling that, $S.E._b = \frac{\sigma_{y \cdot x}}{\Sigma x^2}$ (where $S.E._b$ is the sample estimate of σ_b)

$$S.E._b = \frac{10.048}{\sqrt{4713.76}}$$

$$= .1457$$

$$\text{and } t_o = \frac{.5779}{.1457}$$

$$= 3.97$$

We enter the t-table with $d.f. = N - 2 = 23$ and find the value of $P < .001$.[1]

Therefore, the regression coefficient may be regarded as significantly different from zero.

For the test of the significance of the mean of the dependent variable we test the hypothesis that $\alpha = 0$. For this test we again use the t- criterion,

$$t = \frac{(\bar{Y} - \alpha)\sqrt{N}}{S} \quad \ldots\ldots (8)$$

$$\text{where } S = \sqrt{\frac{\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}}{N - 2}}$$

as in Equation (7).

In our problem:

$$\bar{Y} = 78.76, \qquad N = 25, \qquad s = 10.048$$

$$\text{and } t_O = \frac{(78.76)(\sqrt{25})}{10.048}$$

$$= 39.1$$

[1] Palmer O. Johnson, *op. cit.*

Entering the table of t with $d.f. = 23$ the corresponding probability value, $P < .001$, and hence $\overline{Y}$ is highly significant.

From this analysis, we may conclude then that the prediction equation (4) may be used with confidence in predicting for other samples from the same population. We need, however, a measure of the accuracy of the prediction in each individual case.

The standard error of the estimate Y_E for a specified value of X, say X_O, is given by:

$$s_{Y_E} = \left\{\frac{s^2_Y(1 - r^2)}{N - 2}\left[1 + \frac{(X_O - \bar{X})^2}{s^2_X}\right]\right\}^{\frac{1}{2}} \cdots\cdots\cdots\cdots\cdots\cdots$$

For our problem we have calculated the values of S_{Y_E} for each of the 25 students. These values are recorded in Table 32. We have

TABLE 29. Optimum Estimate Values on Predicting Midquarter Score (Y) from Miller's Analogies (X) with the Corresponding Residuals

Ind.	X	Y	Y_E	$Y_E - Y$	$(Y_E - Y)^2$
1	85	82	87.2193	5.2193	27.2411
2	47	66	65.2591	−0.7409	.5489
3	67	89	76.8171	−12.1829	148.4231
4	75	97	81.4403	−15.5597	242.1043
5	53	86	68.7265	−17.2735	298.3738
6	56	56	70.4602	14.4602	209.0974
7	75	95	81.4403	−13.5597	183.8655
8	54	64	69.3044	5.3044	28.1367
9	43	69	62.9475	−6.0525	36.6328
10	99	103	95.3099	−7.6901	59.1376
11	76	70	82.0182	12.0182	144.4371
12	65	52	75.6613	23.6613	559.8571
13	79	81	83.7519	2.7519	7.5730
14	69	72	77.9729	5.9729	35.6755
15	78	84	83.1740	−0.8260	.6823
16	67	78	76.8171	−1.1829	1.3993
17	75	75	81.4403	6.4403	41.4775
18	73	75	80.2845	5.2845	27.9259
19	72	67	79.7066	12.7066	161.4577
20	51	74	67.5707	−6.4293	41.3359
21	79	86	83.7519	−2.2481	5.0540
22	67	83	76.8171	−6.1829	38.2283
23	74	82	80.8624	−1.1376	1.2941
24	96	98	93.5762	−4.4238	19.5700
25	84	85	86.6414	1.6414	2.6942
Total			1968.9711		2322.2231

also calculated the 98 per cent confidence interval for each student. These confidence intervals enable us to make the statement that a

specified interval will include or cover the true mean midquarter score for all individuals in the population with a specified Analogies score, X_0, and that we may be confident that our statement is correct 98 times out of a hundred.

We may add to the understanding of what the individual estimates yield by estimating the mean midquarter score in two ways: (1) We may estimate the midquarter score of each individual student, and calculate the mean of these estimates, or (2) we may apply the equation (Eq. 4) directly, using the mean value of the Miller Analogy scores for the 25 students.

The calculations involved in Method 1 are given in Table 29. It is found that the mean value of the individual estimates, $\bar{Y}_E$ is

$$\bar{Y}_E = \frac{\Sigma Y_E}{N} = \frac{1968.9711}{25} = 78.758844 \text{ or } 78.76$$

By Method (2) we get

$$\begin{aligned} \bar{Y}_E &= .5779(\bar{X}) + 38.0978 \\ &= (.5779)(70.36) + 38.0978 \\ &= 78.758844 \text{ or } 78.76 \end{aligned}$$

It is noted that the two methods yield exactly the same results carried to 6 decimal places.[1]

As a final step in our analysis we apply the prediction equation (Eq. 4) to another random sample of 25 graduate students chosen from the same student population used to establish the equation. It is good practice to try out the prediction equation on at least one check sample other than the original group. This is done to determine its effectiveness before accepting its validity for application to the general population of which the experimental and check group are random or representative samples.

The necessary calculations are given in Table 30, including the residuals. Comparison between the Y_E's and the Y's reveals close agreement. As we would expect from the principle of least squares, the $\Sigma(Y_E - Y)^2$ (see Table 29) for the sample on which the equation was standardized is smaller than the sum of squares of the residuals for the check sample, that is, $2322.2231 < 3028.8560$ but the difference is not significant:

$$F = \frac{SS_{II}/25}{SS_I/23} = \frac{3028.8560}{2322.2231} \cdot \frac{23}{25} = 1.20$$

$$\text{For } n_1 = 25,\ n_2 = 23,\ P > .05$$

[1] This last analysis has been presented for information. It would not ordinarily be carried out in a research project.

Computation of regression equation from grouped data. Frequently in dealing with regression problems we have a large number of observations. In such cases we would ordinarily group the data. We shall now present an efficient method of setting up the regression equation for grouped data. We first prepare a two way frequency table which consists of a number of square cells, made up of the intersections of the arrays. Those arrays running vertically are called columns and those horizontally are called rows. We may consider the two variates X and Y, X being taken to vary horizontally, i.e., in rows, and Y vertically, i.e., in columns. The class intervals of the X variate constitute the column headings; those of the Y variate, the row headings. An efficient interval size is one which is not greater than one-fourth of the standard deviation. This can be readily estimated from a knowledge of the size of sample and the ratio of the mean range to the standard deviation, which has been established empirically; for example, for the following:

N (*Number in sample*)	$\frac{\textit{Mean Range}}{\textit{Standard Deviation}}$
100	5.02
150	5.30
200	5.49
250	5.63
300	5.76
350	5.85
400	5.94
500	6.07

The range can be estimated from the sample. Then, knowing the size of the sample, it is easy to estimate the standard deviation and take one-fourth of it to obtain the desirable length of interval.

Having set up the appropriate two way table, we proceed to enter each pair of observations as a tally in its appropriate cell.

We shall illustrate the successive steps in the computation by setting up the regression equation for the data in Table 33. This table shows the scores of a random sample of 132 students on two examinations in college biology. One examination was designed to measure the extent to which the student had acquired principles of biology and the other, the acquisition of the ability to use these principles in the solution of problems in biology. We wish to set up the regression equation for the prediction of problem solving ability (Y) from a knowledge of the extent of acquisition of principles (X).

The two-way frequency table (Table 31) was set up with a class

TABLE 30. Estimated Values on Predicting of Midquarter Score (Y) from Miller's Analogies (X) for a Second Random Sample and the Regression Equation for the First Random Sample and the Corresponding Residuals

Ind.	X	Y	Y_E	$Y - Y_E$	$(Y - Y_E)^2$
1	51	73	67.5707	5.4293	29.4773
2	80	96	84.3298	11.6702	136.1936
3	77	85	82.5961	2.4039	5.7787
4	76	93	82.0182	10.9818	120.5999
5	66	85	76.2392	8.7608	76.7516
6	53	77	68.7265	8.2735	68.4508
7	69	54	77.9729	−23.9729	574.6999
8	72	82	79.7066	2.2934	5.2597
9	50	56	66.9928	−10.9928	120.8417
10	76	70	82.0182	−12.0182	144.4371
11	69	86	77.9729	8.0271	64.4343
12	77	85	82.5961	2.4039	5.7787
13	57	86	71.0381	14.9619	223.8585
14	84	97	86.6414	10.3586	107.3006
15	83	99	86.0635	12.9365	167.3530
16	27	46	53.7011	−7.7011	59.3069
17	53	53	68.7265	−15.7265	247.3228
18	75	83	81.4403	1.5597	2.4327
19	48	85	65.8370	19.1630	367.2206
20	69	87	77.9729	9.0271	81.4885
21	53	62	68.7265	−6.7265	45.2458
22	87	87	88.3751	−1.3751	1.8909
23	86	95	87.7972	7.2028	51.8803
24	77	70	82.5961	−12.5961	158.6617
25	92	104	91.2646	12.7354	162.1904
Total					3028.8560

interval of 4 score points for the variable Y, and 4 score points for the variable X. After the scores for each of the 132 students were tallied in their appropriate cells, these tallies were added cell by cell. Then all the numbers in the columns were added to obtain (as indicated in Table 31) the frequency distribution of scores on the informational test. The numbers in the rows were added to obtain the frequency distribution of scores on the problem-solving test.

The regression line whose equation is wanted in this case is the line fitted by the method of least squares to the points whose abscissae are the center of the X-intervals and whose ordinates are the means of the corresponding distributions of Y in the columns centered at the appropriate X's (see Table 31). In calculating the

TABLE 31. Calculation of the Regression Coefficient from Grouped Data

SCORES ON INFORMATION TEST (X)

SCORES ON APPLICATION TEST (Y)	39.5-43.5	43.5-47.5	47.5-51.5	51.5-55.5	55.5-59.5	59.5-63.5	63.5-67.5	67.5-71.5	71.5-75.5	75.5-79.5	79.5-83.5	83.5-87.5	87.5-91.5	91.5-95.5	95.5-99.5	99.5-103.5	103.5-107.5	107.5-111.5	111.5-115.5	115.5-119.5	119.5-123.5	123.5-127.5	F	y	Fy	Fy^2	Σfx	yΣfx
57.5 - 59.5																	1^{5}						1	10	10	100	5	50
55.5 - 57.5																1^{4}	1^{5}	1^{6}					3	9	27	243	15	135
53.5 - 55.5											1^{-1}							1^{6}	1^{7}				3	8	24	192	12	96
51.5 - 53.5											2^{-2}			1^{2}								1^{10}	4	7	28	196	10	70
49.5 - 51.5								1^{-4}			1^{-1}	1^{0}	2^{2}					1^{6}		1^{8}			7	6	42	252	11	66
47.5 - 49.5										2^{-4}	1^{-1}	1^{0}	3^{3}	1^{2}	1^{3}	2^{8}	2^{10}						13	5	65	325	21	105
45.5 - 47.5									1^{-3}			1^{0}	1^{1}	3^{6}	1^{3}		1^{5}						8	4	32	128	12	48
43.5 - 45.5								2^{-8}			1^{-1}		1^{1}	1^{2}	2^{6}								7	3	21	63	0	0
41.5 - 43.5				1^{-8}	1^{-7}			2^{-8}			1^{-1}		1^{1}	2^{4}	1^{3}								9	2	18	36	-16	-32
39.5 - 41.5	1^{-11}						1^{-5}	3^{-12}	3^{-9}	2^{-4}	2^{-2}	3^{0}	1^{1}	1^{2}	1^{3}		1^{5}			$\bar{Y} = 40.5$			19	1	19	19	-32	-32
37.5 - 39.5				2^{-16}		1^{-6}	1^{-5}			1^{-2}	3^{-3}	2^{0}					1^{5}						11	0	0		-27	0
35.5 - 37.5		1^{-10}		1^{-8}	2^{-14}		1^{-5}			2^{-4}			1^{1}		1^{3}								9	-1	-9	9	-37	37
33.5 - 35.5			1^{-9}	1^{-8}	2^{-14}	2^{-12}	2^{-10}	1^{-4}	2^{-6}		1^{-1}	1^{0}	1^{1}										14	-2	-28	56	-63	126
31.5 - 33.5				1^{-8}			1^{-5}		1^{-3}	1^{-2}			1^{1}										5	-3	-15	45	-17	51
29.5 - 31.5			1^{-9}			2^{-12}	1^{-5}	4^{-16}	2^{-6}														10	-4	-40	160	-48	192
27.5 - 29.5								1^{-4}		1^{-2}													2	-5	-10	50	-6	30
25.5 - 27.5	1^{-11}			1^{-8}			1^{-5}																3	-6	-18	108	-24	144
23.5 - 25.5							1^{-5}																1	-7	-7	49	-5	35
21.5 - 23.5			1^{-9}																				1	-8	-8	64	-9	72
19.5 - 21.5								1^{-4}															1	-9	-9	81	-4	36
17.5 - 19.5						1^{-6}																	1	-10	-10	100	-6	60
(1) F	2	1	3	7	5	6	9	15	9	9	13	9	12	9	7	3	7	3	1	1	0	1	132 (N)		132 (ΣFy)	2276 (ΣFy^2)	-208	1289 (ΣFxy)
(2) x	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10						
(3) Fx	-22	-10	-27	-56	-35	-36	-45	-60	-27	-18	-13	0	12	18	21	12	35	18	7	8	0	10	-208 (ΣFx)					
(4) Fx^2	242	100	243	448	245	216	225	240	81	36	13	0	12	36	63	48	175	108	49	64	0	100	2744 (ΣFx^2)					

TABLE 32. Standard Errors of Estimated Values of Y for Different Values of X_o with Corresponding 98 Per Cent Confidence Intervals

(1) N	(2) X_o	(3) $(X_o - 70.36)^2$	(4) $(.0205)(X_o - 70.36)^2$	(5) $S_{Y_E}^2$ Col. (4) + 4.026	(6) S_{Y_E}	(7) $t_{.02} \cdot s_{Y_E}$	(8) *Confidence Interval*
1	85	214.3296	4.3938	8.4198	2.901	7.25	77.75 – 92.25
2	47	545.6896	11.1866	15.2126	3.901	9.75	37.25 – 56.75
3	67	11,2896	.2314	4.2340	2.055	5.14	61.86 – 72.14
4	75	21.5296	.4414	4.4674	2.113	5.29	69.71 – 80.29
5	53	301.3696	6.1781	10.2041	3.194	7.99	45.01 – 60.99
6	56	206.2096	4.2273	8.2533	2.872	7.18	48.82 – 63.18
7	75	21.5296	.4414	4.4674	2.055	5.14	69.86 – 80.14
8	54	267.6496	5.4868	9.5128	3.025	7.56	46.44 – 61.56
9	43	748.5696	15.3457	19.3717	4.401	11.00	51.95 – 73.95
10	99	820.2496	16.8151	20.8411	4.565	11.41	87.59 – 110.41
11	76	31.8096	.6521	4.6781	2.162	5.41	70.59 – 81.41
12	65	28.7296	.5890	4.6150	2.148	5.37	59.63 – 70.37
13	79	74.6496	1.5303	5.5563	2.357	5.89	73.11 – 84.89
14	69	1.8496	.0379	4.0639	2.015	5.04	63.96 – 74.04
15	78	58.3696	1.1966	5.2226	2.285	5.71	72.29 – 83.71
16	67	11.2896	.2314	4.2574	2.063	5.16	61.84 – 72.16
17	75	21.5296	.4414	4.4674	2.055	5.14	69.86 – 80.14
18	73	6.9696	.1429	4.1689	2.041	5.10	67.90 – 81.1
19	72	2.6896	.0551	4.0811	2.020	5.05	66.95 – 77.05
20	51	374.8096	7.6836	11.7096	3.421	8.55	42.45 – 59.55
21	79	74.6496	1.5303	5.5563	2.357	5.89	73.11 – 84.89
22	67	11.2896	.2314	4.2574	2.063	5.16	61.84 – 72.16
23	74	13.2496	.2716	4.2976	2.073	5.18	68.82 – 79.18
24	96	657.4096	13.4769	17.5029	4.183	10.46	85.54 – 106.46
25	84	186.0496	3.8140	7.8400	2.800	7.00	77.00 – 91.00

$$s_{Y_E}^2 = \frac{s_y^2(1 - r^2)}{s_x^2(N - 2)}[s_X^2 + (X_O - \bar{X})^2]$$

$$s_{Y_E}^2 = \frac{(162.235)(1 - .404)}{(196.406)(23)}[196.406 + (X_O - 70.36)^2]$$

$$s_{Y_E}^2 = .0205[196.406 + (X_O - 70.36)^2]$$

$$s_{Y_E}^2 = 4.026 + .0205(X_O - 70.36)^2$$

slope of the regression line for grouped data we make use of a computation variable. The computation variable is given in Table 31, in the row (2) headed by x and in the column (2) by y. The values, x's, are secured as follows: take any class interval of the X-variate as the position of the assumed mean, say $\overline{X}_A$. In our example, we have taken $\overline{X}_A$ as the midpoint of the class interval 83.5 – 87.5, or the point 85.5. Then we subtract the value of $\overline{X}_A$ from each of the midpoints of the class-intervals in turn and divide the difference by the length of the class interval, h, or 4. This gives us the values shown in the row (2), for example, in the interval 39.5 – 43.5 we have

$$\frac{41.5 - 85.5}{4} = -11$$

and for the interval 95.5 − 99.5

$$\frac{97.5 - 85.5}{4} = 3$$

It is apparent that in this way the values of the computation variable are reduced to the smallest convenient size. They may be written down directly without actually calculating each one. Opposite the midpoint of the interval chosen as the assumed mean, we put 0, and proceed to write in the other values as shown. Likewise the computation variable *y's* are written as in the colmun *y*.

The slope of the regression equation, b_{yx}, in terms of the computation variable and the adjustment made to secure the value in raw score form is

$$b_{yx} = \frac{\Sigma Fxy - \dfrac{(\Sigma Fx)(\Sigma Fy)}{N}}{\Sigma Fx^2 - \dfrac{(\Sigma Fx)^2}{N}} \cdot \frac{(h)(k)}{(h)^2} \qquad (10)$$

where h = length of the class-interval for the X-scores and k = length of the class-interval for the Y-scores.

The values needed for equation (10) are given in the appropriate columns at the right of the table and the appropriate rows in the lower part of the table, Table 31.

Only the method of obtaining the values in column (5) needs further explanation. To obtain the values given in column (5) we multiply the frequencies in each row of Table 31 by the corresponding value of the computation variable x; write the products in the upper right-hand corners of the respective cells; and add these values for each row. For the first row of table we multiply the frequencies by 5 (since $x = 5$ for this column) and obtain the value shown; for the second row we have 1×4, 1×5, 1×6, the sum of which is 15, which is the second value under column (5). In the same way all the values under column (5) are obtained. The values in column (6) are the products of the values in columns (2) and (5).

For our problem we have

$$N = 132 \qquad \Sigma Fxy = 1289$$
$$\Sigma Fx = -208 \qquad h = 4$$
$$\Sigma Fx^2 = 2744 \qquad k = 2$$
$$\Sigma Fy = 132$$

The student should make it habitual to check the accuracy of all the calculations. Check the above, for instance.

TABLE 33. Scores on Biology Test for a Random Sample of 132 Students: Scores on Information Test X, on Application Test Y

X	Y	X	Y	X	Y
63	34	83	44	90	49
71	42	80	52	69	31
70	41	89	49	52	37
119	50	98	44	40	41
109	57	73	35	82	40
75	30	65	30	90	37
88	33	62	30	108	54
83	55	114	54	83	40
68	20	105	39	98	37
59	35	88	35	61	18
55	43	78	49	80	39
106	47	69	51	70	40
56	35	67	36	60	30
81	51	79	29	66	34
102	48	80	38	71	31
94	43	47	36	85	46
97	40	68	42	43	26
84	39	93	44	65	32
91	51	78	37	53	35
85	41	51	34	88	45
106	49	92	46	68	41
86	49	76	36	93	46
104	41	105	57	91	47
78	40	55	32	101	56
91	51	86	50	94	40
82	43	71	30	91	41
64	34	70	31	73	33
55	38	68	28	99	47
87	40	81	39	99	45
50	30	81	48	66	40
75	46	65	39	78	40
73	41	104	49	56	37
59	43	88	43	93	48
91	48	78	32	85	38
80	52	84	40	58	36
105	59	92	47	92	43
97	48	84	35	75	31
77	39	78	48	66	27
124	52	66	25	69	44
68	34	94	53	111	50
101	49	52	39	73	35
81	34	61	38	73	41
69	44	96	43		
73	40	53	27		
63	34	49	22		

Substituting these values in equation (10) we have

$$b_{yx} = \frac{1289 - \dfrac{(-208)(132)}{132}}{2744 - \dfrac{(-208)^2}{132}} \cdot \frac{(4)(2)}{(4)^2}$$

$$= .309778$$

$$\bar{X} = 85.5 + \left(\frac{-208}{132}\right)4 = 79.1968$$

$$\bar{Y} = 38.5 + \left(\frac{132}{132}\right)2 = 40.5$$

The regression equation of Y on X in raw score form is

$$Y_E = \bar{Y} + b_{yx}(X - \bar{X})$$

For our problem

$$\begin{aligned} Y_E &= 40.5 + .309778\,(X - 79.1968) \\ &= .3098\,X + 15.9648 \qquad (11) \end{aligned}$$

We now test the significance of the regression coefficient, .3098: $t_0 = 4.98$ and $P < .001$ for $d.f. = 130$. Therefore, there is a significant regression of Y on X. Likewise the test of significance of $\bar{Y} = 40.5$ gives $t_0 = 152.11$ and $P < .001$ for $d.f. = 130$; accordingly, it may be accepted as significant.

Up to this point we have shown how to set up the equation for predicting one variable from another under the conditions that a linear relationship exists betwen the two variables. When problems are encountered where the relation between the two variables is curvilinear, other methods of analysis become necessary. The discussion of these is beyond the scope of this book.

If we have measures of several traits which may be useful in predicting an unknown characteristic, methods involving the multivariate case are employed. Here also the regression may be linear or nonlinear. The multivariate linear regression will be discussed in the materials to follow. The problem is essentially that of determining the weightings to be assigned to the respective independent variates to bring about the best prediction of the dependent variate.

Multiple regression. We shall describe next the mathematical model for the multiple-regression problem.

In multiple-regression analysis we assume that we have a single dependent variate Y and several independent fixed variates X_i, say p in number. These variates are connected by a mathematical model defined as follows:

$$Y_E = b_1X_1 + b_2X_2 + \cdots\cdots + b_pX_p + e \qquad (1)$$

where the b_i are estimates of true regression coefficients β_i, and the residual e is the estimate of a true residual assumed to be normally and independently distributed with mean o and variance σ^2. Since the independent variates, X_i, are assumed to be fixed values, Y_E is normally and independently distributed with mean $\sum_{i=1}^{p} \beta_i X_i$ and variance σ^2. A third assumption is that the independent effects are additive.

The statistician's problem then is to determine, with the help of n observations on each of the variates $X_1, \ldots, X_p$ and Y, the most suitable coefficients $b_1, \ldots, b_p$ in such a prediction formula as (1), which is to be used with future values of $X_1, \ldots, X_p$ to get estimates Y_E of the values Y to be associated with these future X_i.

The method of least squares, which has certain desirable properties, is used to determine the coefficients b_i, called the partial regression coefficients. By this method the sum of the squared deviations of the observed Y's from the corresponding values of the Y_E's, calculated from the prediction formula (1) by substituting in it the observed X's, is made a minimum.

In judging the accuracy of the coefficients, b_i, and the over-all goodness of fit of the regression model, use is made of the minimum value obtained from the sum of squares of the residuals, the e's, or, the deviations, $(Y_E - Y)$'s. This is done partly through the standard errors of the b_i's and of the forecast Y_E, and partly through the correlation coefficient, known as the multiple correlation, R, between the Y's and the Y_E's for the observed sample. For the multiple R there is available exact interpretation in terms of probability.

The application of the methods of least squares leads to p, minimizing equations for the b's:

$$\begin{aligned}
b_1\Sigma x_1^2 + b_2\Sigma x_1x_2 + \cdots + b_p\Sigma x_1x_p &= \Sigma x_1y \\
b_1\Sigma x_1x_2 + b_2\Sigma x_2^2 + \cdots + b_p\Sigma x_2x_p &= \Sigma x_2y \\
&\vdots \\
b_1\Sigma x_1x_p + b_2\Sigma x_2x_p + \cdots + b_p\Sigma x_p^2 &= \Sigma x_py
\end{aligned}$$

These are called the normal equations for the b_i's. We have p equations in the p unknowns, b_i.

If these equations are solved by the matrix inversion method, which utilizes either Fisher's c-values or the k-values,[1] there is available at once the necessary tools to determine an estimate of σ^2 and the variance of each b_i, as well as the goodness of fit of the entire regression model. For the short-cut matrix-inversion technique, some one of the variants of the Doolittle method described by Dwyer[2] provides the most useful method for the research worker. The matrix-inversion technique furnishes all the data with which to perform the necessary tests of significance and also for setting up confidence intervals for the various parameter values. For the complete solution of the matrix-inversion technique by the Doolittle method, the reader is referred to Johnson.[3]

The important question with respect to the regression coefficients and the predicted values is the accuracy that may be expected upon application of the prediction formula established upon one sample to a new sample. The correlation of the Y's in a new sample with the Y_E's, calculated from the prediction formula that is, the multiple correlation coefficient obtained on the new sample will usually be less than the original R, to an extent to be expected on the basis of mathematical theory. It has been established that the formula which gives maximum prediction efficiency for a new sample is that which was determined by the method of least squares from the original sample.

The choice of a criterion often introduces difficult problems.[4] For example, we may wish to measure "teaching ability" for the purpose of developing an efficient forecasting formula for the guidance and selection of young people who might be successful as teachers. In order to do so we must develop a measure of "teaching ability." There will be several criteria available, all appearing to be more or less roughly correlated with the thing which we have in mind to measure, and also with one another. If the coefficients of correlation with the fundamental thing whose measure is under search could be accurately estimated, it would be possible to compute by least squares a set of partial regression coefficients or weights to be applied to the various measures so as to

[1] In education and in certain other fields, it is usually of interest to obtain the zero-order correlation coefficients, and the Beta or standard partial regression coefficients. Therefore, the correlation matrix rather than the product-sums matrix is used (see Johnson).

[2] P. S. Dwyer, "The solution of simultaneous equations," *Psychometrika*, 1941, 6:101–129.

[3] Palmer O. Johnson, *Statistical Methods in Research* (New York: Prentice-Hall, Inc., 1949).

[4] Harold Hotelling, "Problems in prediction," *The American Journal of Sociology*, 1942, 48:61–76.

obtain that index correlating most highly with the real thing. There is, however, another consideration. Not only is it desired to estimate the real thing by means of the best combination of proposed criteria say, the Y's, but it is also desired to predict the real thing by means of some other variates say X's, belonging to a different group than the group to which the several proposed criteria belong. In the example, the proposed criteria, the Y's, would be observations of various kinds on teachers while the X's would be observations on persons contemplating teaching. The ordinary procedure in such a case is for the investigator to begin by choosing an index, say Y', chosen according to his judgment or the consensus of several judgments. This is done without any definite reference to the other set of variates, the X's. It is after Y' has been selected that the later step is taken of computing the regression coefficients of the X's by the principle of least squares in order to estimate the Y''s as precisely as possible.

Hotelling has suggested a method which takes into account the predictors, X's, in the process of selecting the weights that determine the predicted Y'. In this approach it is assumed that some of the criteria, $y_1, y_2, \ldots$, can be predicted from the X's more precisely than others. His contention is that these Y's should be regarded as of greater importance in making up the index Y'. Thus, for example, if say, y_1 and y_2 are correlated equally closely with the real thing, say Z, that is sought, but if y_1 can be predicted more accurately from the X's than can y_2, then greater success will be achieved in predicting Z by predicting y_1 than by predicting y_2. Still greater success would be attained by using certain linear combinations of y_1 and y_2. In view of the fact that the correlation of the Y's with the X's and with one another can be obtained from objective observations, Hotelling in 1933 [1] suggested that the most valuable Y' will frequently be the one that can be predicted most accurately and indicated a technique yielding the most predictable criterion. Later he made a more detailed study of this problem.[2]

Other investigators have contributed to the discovery of the most predictable criterion and of other linear functions of the criteria $y_1, y_2 \ldots$, which possess properties that make them of value whenever a criterion other than the most predictable one is of interest.

The objection has been raised that the choice of a purpose and

[1] Harold Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, 1933, 24:417–441.

[2] Harold Hotelling, "Relation between two sets of variates," *Biometrika*, 1936, 28:321–379.

therefore the specification of the variate to be predicted must be regarded as outside the statistical theory of prediction.[1] Hotelling points out that this emphasis upon definition seems to require that all analysis be based on definite measurements of the true criterion rather than on its symptoms or evidences. He points out that we can measure only various aspects of behavior which may be regarded as the manifestations of the trait under prediction. That is, if we are to predict the real criterion as a consequence of certain predisposing circumstances, and to check the accuracy of the prediction by observations, it is necessary to use as a criterion of the trait some function of observable variates only. The choice as to which of the many possible functions that should be used is to some extent arbitrary. However, the procedure obviously will be to select some function which intuitively seems to be correlated with what we wish to measure but cannot. Within the class of reasonably suitable functions, it is desirable to select one that promises to be predicted well rather than one capable of being only predicted poorly or not at all.

The nature of multiple regression can probably be best understood by studying an illustration for this purpose; the practical situation of predicting the scholastic success of a sample of students by combining the scores on three different predictor measures has been chosen.[2] The investigation to be reported is much more comprehensive than can be described here, but it will serve to illustrate the solution of the problem of setting up a prediction equation in the multivariate case. It will also illustrate how the findings of the investigation were made available for use of the several advisers of students. An exhaustive investigation of twelve different independent variables or predictors had resulted in determining three that were statistically significant and of stability established by trial on samples of students not included in the original sample upon whom the equation was determined. Although prediction equations were set up for each of the divisions of the College of Agriculture, Forestry, and Home Economics, of the University of Minnesota the illustration that follows is for a representative sample of 119 students from the Division of Home Economics.

The criterion variable, Y, was the first year honor-point ratio. The predictors were (1) the high school percentile rank, converted to standard deviation units called probits (X_1); (2) score on the

[1] Paul Horst *et al*, "The prediction of personal adjustment," *Social Science Research Council Bulletin*, 1941, 48.

[2] Edward M. Freeman and Palmer O. Johnson, "Prediction of success in the college of agriculture, forestry, and home economics," pp. 33–65 in *University of Minnesota Studies in Predicting, Part I* (Minneapolis: University of Minnesota Press, 1942).

Johnson Science Application Test (X_2); and score on the Co-operative Algebra Test, (X_3).

The predictive formula, or multiple regression equation based on the complete observational data of the sample of 119 students was as follows:

$$Y_E = .7393 + .01646\, X_1 + .01170\, X_2 + .00586\, X_3 \qquad (2)$$

where Y_E is the predicted honor point ratio at the end of the freshman year. The coefficients of X_1, X_2, and X_3 are the partial regression coefficients. Each regression coefficient indicates the average change to be expected in the dependent variable, the criterion Y_E, for a unit change in the particular independent variable, with the other independent variables assumed to be held constant. For example, .01646, the coefficient of X_1, measures the added honor point ratio expected for an addition of one unit change in high school percentile rank (in probits), while leaving the amounts of the other variates, score on Science test, and score on Algebra test, unchanged.

The multiple correlation of $Y_{Y.123}$ was found to be .74.

For practical purposes, equation (2) was converted to the following form by transferring the constant (.7393) to the other side of the equation and then dividing the whole equation through by the partial regression weight for the Co-operative Algebra Test (.00586):

$$W = 2.81\, X_1 + 2.00\, X_2 + X_3 \qquad (3)$$

where W is the predicted measure of first-year achievement in the Division of Home Economics.

A probability or expectancy table was prepared from the bivariate frequency distribution of earned first-year honor-point ratios and first-year honor point ratios predicted from equation (2) (see Table 34). The different grade levels are set across the top of the table; the predicted honor point ratio (Y_E) intervals are placed along the right hand side of the table; the predicted score (W) intervals are at the left-hand side. The table may therefore be entered from either side, depending upon the form of the multiple-regression equation used, to obtain the probability of a student's earning a grade equal to or above a specified level.

If information is available, therefore, concerning an applicant's high school percentile rank, his score on the Johnson Science Application Test, and his score on the Co-operative Algebra Test, these quantities may be substituted for the respective X's in the predictive formula, and then, by entering the probability table, the

chances of the applicant's earning a grade equal to or above a certain level may be obtained. For example, suppose an applicant to the Division of Home Economics had a high school percentile rank of 70, a score of 32 on the Johnson Science Application Test, and a score of 21 on the Co-operative Algebra Test. When Equation (2) is used the predicted honor-point ratio (Y_E) would be .91 for the applicant, or when Equation (3) is used the predicted score (W) would be 282. Entering Table 34 with either Y_E or the W value we find that there are 53 chances in 100 of the applicant's making an average grade of E or better, 50 chances in 100 of his making an

TABLE 34. Probability Table Giving the Chances in 100 that a Freshman in the Division of Home Economics with a Particular Predicted Score (Y_E or W) Will Earn a Grade Equal to or Above Different Specified Grade Levels

Predicted Score (W)	*Chances in 100 of Earning a Grade Equal to or Above* E	D	C	B	*Predicted Honor Point Ratio* (Y_E)
Below 211	17	12	2	0	Below .50
211–253	31	27	9	0	.50– .74
254–296	53	50	29	0	.75– .99
297–338	72	71	53	0	1.00–1.24
339–381	87	87	76	40	1.25–1.49
382–424	97	96	93	70	1.50–1.74
425–466	98	98	96	80	1.75–1.99
467 and above	100	100	100	100	2.00 and above

average grade of D or better, 29 chances in 100 of his earning an average grade of C or better, and 0 chances in 100 of his earning an average grade of B or better during his freshman year.

All of this information is now placed each year in the hands of every freshman adviser to be used for guidance purposes.

The discriminant function. When the predicted criterion is discrete, we classify our observations into broad groups, for example, "good" and "bad," "male" and "female," "industrial arts," and "college preparatory" curriculum, without any fine gradations of the thing that we try to predict. However, we may have continuous variates, such as age, I.Q., stature, mechanical ability, on the basis of which we are to select individuals to be placed in one group or the other. Given any set of data of this kind, it is possible to determine a function of the measured continuous variates alone by which individuals of the sample can be classified into the two classes with error which in a certain reasonable sense may be said to be minimum. Such a function was proposed by R. A. Fisher,

who called it the discriminant function. The problem to be solved is the determination of that linear combination of the various measurements which will best discriminate between the two groups. For illustration of its method of calculation and the appropriate tests of significance the reader is referred to Johnson.[1]

PROBLEMS AND PROCEDURES INVOLVED IN THE USE OF MULTIVARIATE ANALYSIS

The problems of predicting events and of estimating structural relationships in a multivariate situation are of wide occurrence and importance. They include many problems, such as those arising from the adjustment of individuals to school, occupation, marriage, and the law. The factors operative in school, in vocation, and in marital adjustment are numerous. Complex problems of this sort can be solved only through a multivariate approach. The educator, the psychologist, the sociologist, and the criminologist, therefore require an understanding of the means of the control of situations through the prediction of subsequent events, outcomes, and interrelationships.

Our interest here is chiefly in the vocational and educational fields. In the former, considerable work has been done in the attempt to predict, prior to employment, the chances of success of an individual (1) in a single job, (2) in some one or more of a number of jobs, and (3) in determining to which one of a number of available jobs an individual might be assigned with the least probability of a misclassification. Barr and his students have made intensive studies of the factors associated with teaching success.[2] In the field of education much attention has been given to the study of children in varying stages of their educational career with a view to predicting the probable success in subsequent stages.[3]

More critical examination of these problems has led to the study of differential prediction aimed at answering the question as to which field an individual is most likely to succeed in. The more immediate problem is that of predicting the relative probabilities of success in particular types of subject matter areas, such as mathe-

[1] Palmer O. Johnson, *Statistical Methods in Research* (New York: Prentice-Hall, Inc., 1949).

[2] A. S. Barr *et al.* "The prediction of teaching efficiency," *The Journal of Experimental Education*, 1946, 15.

[3] Edward M. Freeman and Palmer O. Johnson, "Prediction of success in the college of agriculture, forestry, and home economics," pp. 33–65 in *University of Minnesota Studies in Predicting, Part I* (Minneapolis: University of Minnesota Press, 1942). See also Noel Keys, "The value of group test I.Q.'s for prediction of progress beyond high schools," *Journal of Educational Psychology*, 1940, 31:81–93.

matics, languages, or the sciences, which are related in turn to the training programs for different occupations. Continuous emphasis is being given in modern education to the need of flexibility in courses of study and in educational programs with a view to better adaptation to individual differences. For the effectiveness of these plans early discovery of the abilities, interests, personalities, and deficiencies of children and their significance to continuous growth and development make prediction studies of fundamental importance.

The prediction of the number of pupils to be expected in the public schools on the basis of birth rate trends, the estimation of the necessary number of teaching personnel, and the number of school buildings to be provided—these and other imminent problems force the recognition of the need for continuous collection of basic facts for an intelligent program of action.

It is not possible here to outline all the designs of studies needed for attacking the many complex problems of prediction mentioned in this chapter. For further information the student should read the reports of investigations, cited in the references, giving particular attention to the plan of investigation, the method of analysis, and the presuppositions underlying the validity of the study.

The principal steps in the design of prediction studies which have been enumerated in this chapter may be summarized as follows:

1) Specification of the purpose of the study including a clear statement of underlying assumptions or presuppositions must be made. The specification of the domain of study should include the definition of the population about which the inquiry is planned to supply numerical information of known precision.

2) The method of selecting the sample, including the choice of the sampling unit, the type of sampling adopted, the size of the sample, the proportion it forms of the population covered should be indicated. The sample will supply both objective estimates of the functions under evaluation and information for estimating the precision of predictions.

3) Definition of the criterion or criteria, that is, the measure(s) of "success" in the area of human activity under prediction must be developed; such, for example, as scholastic grade or honor-point ratios, for academic achievement; salary increases or promotions for vocational performance; reputed happiness for marital adjustment; good behavior on release from confinement for success on parole. The criterion measure must have relevance to the ultimate criterion and be at least significantly reliable. For further aid, see

discussion of "The Criterion," by Rulon.[1] For the discussion of criteria, in differential prediction, see Tucker (pp. 62–70), Bennett and Seashore (pp. 71–79), and Dyer (pp. 80–87) in *Exploring Individual Differences.*[2]

4) A precise determination must be made of the data to be collected. The data required depend on their contribution to the prediction of the criteria, e.g., background factors, abilities, aptitudes, personality characteristics, biographical inventory, performance tests, other experimental measures designed for the study, standardized methods of observation, projective techniques, and others. In general, the contribution of any measure to the prediction of the criterion is a function of the relation of the measure to the criterion and to the other measures.

5) A selection must be made of the tests and other measurements to be administered to the sample of experimental subjects. A combination of these measures will be employed to predict the criterion. The technique of multiple regression is used to select the best measures and to allot to them the best weights. Where the criterion is a dichotomy or multiple classification, discriminatory function analysis may be employed. The complete solution of the problem of prediction by the methods of multivariate analysis involving as it does tests of significance and problems of estimation has been found extremely useful.[3]

6) A try-out of the measures found valid in operation (5) should be made on at least another check sample of individuals not involved in the original analysis for the purpose of noting the stability and validity of the relation found between the combination of the predictor factors and the criterion in the original sample.

7) If operation (6) confirms the expectations of the relationship found in (5) to be sufficiently high for practical results the battery of tested measures may then be applied to the general population or to other representative samples of it.

SUMMARY

The methods of this chapter treat a wide class of statistics known as regression coefficients. The idea of regression is traditionally in-

[1] Educational Testing Service, "Validity, norms, and the verbal factor," *Proceedings of the 1948 Invitational Conference on Testing Problems* (New York: 1948).

[2] American Council on Education, *Exploring Individual Differences, A Report of the 1947 Invitational Conference on Testing Problems* (Washington, D. C.: 1948).

[3] Robert Jackson, "The selection of students for freshman chemistry by means of discriminant functions," *Journal of Experimental Education,* 1950, 18:209–214, and Palmer O. Johnson, *Statistical Methods in Research* (New York: Prentice-Hall, Inc., 1949).

troduced as a subsidiary problem under correlation theory. Regression is, however, a more general and a simpler idea than correlation. The regression coefficient is also of wide interest and great scientific importance. The methods employed are based on the principle of least squares. These are essential in testing significance and in obtaining standard errors of the various estimates specified.

In the simplest case of prediction we use the information on one characteristic or factor, mathematically known as the independent variable, for the purpose of estimating or predicting a second factor, the dependent variable. In the most practical case, that of multiple regression, we combine the information available in a number of characteristics into a single system for the purpose of estimating or predicting a criterion, the dependent variable.

If the criterion consists of mutually exclusive classifications or categories we use the discriminant function. Through its use we may classify an individual on whom we have measures on the several independent variables into his appropriate class.

The chapter provides first a quantitative discussion designed to familiarize the reader with the basic ideas underlying simple and multiple regression. In the case of simple regression, the most efficient process of calculating the values needed to set up the equation for estimation or prediction is presented for both grouped and ungrouped data.

CHAPTER X

Correlation Analysis

CORRELATION AS AN EXTENSION OF REGRESSION THEORY

In this chapter we are concerned with problems in which the investigator is primarily interested in the measurement of the magnitude of relationship existing between two variables rather than in the use of a relationship for predicting one variable from a knowledge of another. Measurement of the intensity of relation between two variables is an extension of the theory of linear regression, just treated. In this form the analysis is included under correlation theory.

It will be recalled that in an analysis of the problem of regression involving grouped data a least-square regression line was fitted to a set of points whose abscissae were the centers of the X-intervals and whose ordinates were the means of the corresponding distributions of Y in the columns centered at the appropriate X's. Similarly, another least-square regression line may be fitted to another set of points, the means of the X-distribution in the rows plotted against the center of the corresponding Y-intervals. Drawn on the same diagram with the two axes, say OX and OY at right angles, the two regression lines intersect at a point corresponding to the means of the X- and Y-distributions. Now, in the condition for perfect association, i.e., when X is uniquely determined by Y (and *vice versa*), all the frequencies occur in cells about one diagonal in the correlation table. A negative correlation gives the same appearance of clustering as a positive one, but the points follow a different diagonal. When there is a perfect relationship between the variables X and Y, the two regression lines coincide. When there is no relationship between the variables the two regression lines be-

A. Ages of Husbands and Wives. N = 5,317. (Data from Yule, G. Udny, *Theory of Statistics* [1929] p. 159.)
B. Verbal Score and Binet Intelligence Quotient. N = 500. (*The Intelligence of Scottish Children*, University of London Press, 1933, p. 96.)
C. Stature of Sons and Fathers. N = 1078. (*Biometrika* 2 [1903], p. 415.)
D. Final Scores and Initial Scores. N = 282. (Johnson, P. O. Unpublished.)
E. Daughters' Children and Mothers' Children. N = 1,000. (*Phil. Trans.* A. vol, cxcii [1899] table IV.)
F. Reaction to Sight and Head Length. N = 4,690. (*Biometrika*, 18 [1926], p. 207.)

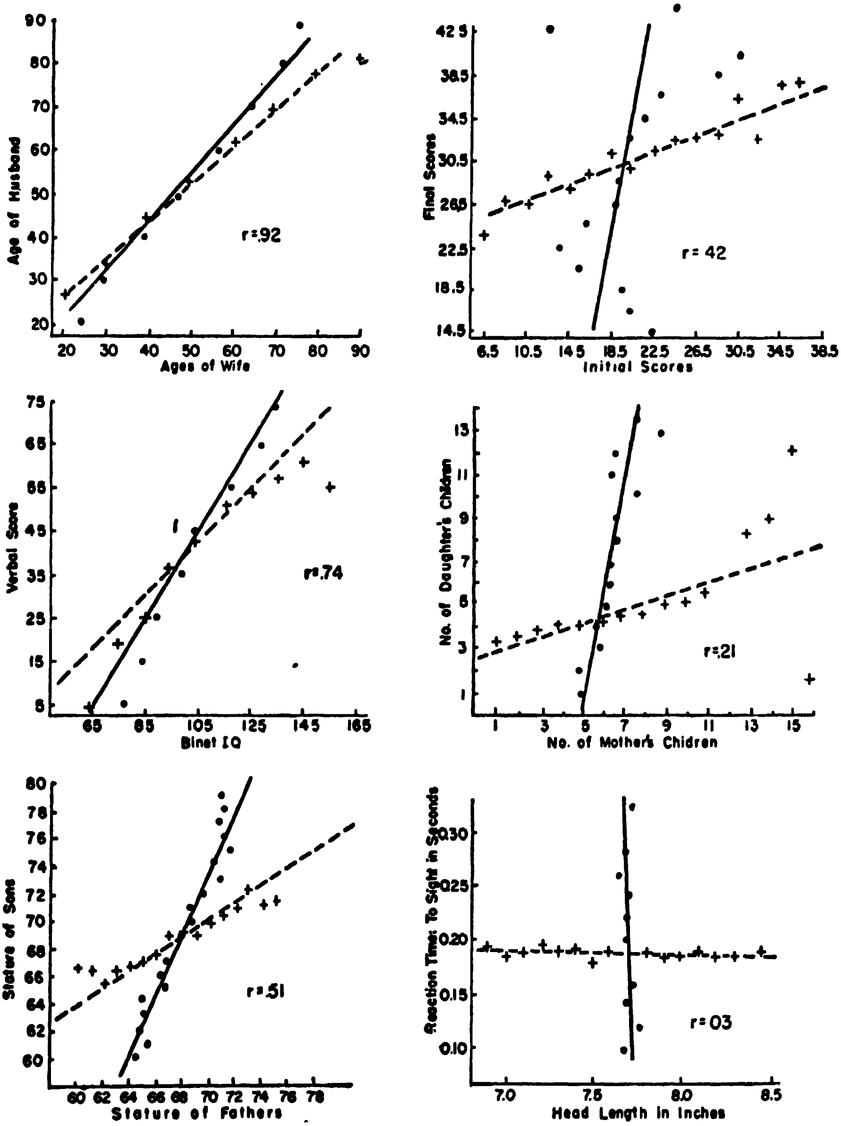


FIGURE 5. The broken lines are the regression of Y on X; the solid lines are the regression of X on Y; and the dots are the means of horizontal arrays. The xx indicate the means of the vertical arrays.

come parallel to the respective axes and perpendicular to one another.

Varying strengths of relationship between perfect relationship and no relationship are indicated by the varying divergence of the two regression lines. These facts are illustrated by the set of six regression diagrams in Figure 5. It is observed that the stronger the relationship between X and Y, the greater is r and the closer together the two regression lines. This observation also suggests that since the angle of divergence between the two regression lines becomes greater with decrease in relationship between the two variables, this fact may be used in obtaining a quantitative index of how much the variables are correlated. We now proceed to explore this problem and to translate the geometrically expressed ideas into analytical language.

A first need arises of rendering the slopes of the regression lines independent of the units of measurement used. The consequence of this dependence is noted if, for example, all the values of the X-variate were multiplied by three and those of the Y-variate by six. The slopes of the regression line would be changed. Such an operation would not alter the degree of correlation between X and and Y, since changes in the scale of measurement for either or both sets of measurements do not change the degree of relationship. The transformation required to render the original measures of the two variates independent of scale is to convert all the original measures to standard measure. If this is done both variables will have a standard deviation of unity. The slopes of the regression lines and their angular separation then become independent of the units of measurement originally employed by fitting the lines to the means of standard scores of the respective columns and rows.

With the expression of the two variables in standard measure the slopes of the regression lines are equal and also equal to the correlation coefficient. The correlation coefficient when calculated may then be used directly to predict standard scores, Z_x's, from a knowledge of standard scores, Z_Y's, and vice versa by the following equations of the two regression lines:

$$\bar{Z}_x = rZ_y \quad \text{(1)}$$
$$\bar{Z}_y = rZ_x \quad \text{(2)}$$

where $\bar{Z}_X$ and $\bar{Z}_Y$ are estimated standard scores, Z_X and Z_Y are standard scores.

The relation between the regression coefficients used in predicting deviation scores and the coefficient of correlation is given by

$$b_{yx} = r\frac{s_y}{s_x} \quad \text{(3)}$$

$$b_{xy} = r\frac{s_x}{s_y} \quad \text{(4)}$$

where b_{yx} and b_{xy} are the regression coefficients of Y on X, and X on Y, respectively; and s_y and s_x are the standard deviations of Y and X, respectively.

By multiplying equations (3) and (4) we obtain

$$r^2 = b_{yx} \cdot b_{xy} \quad \text{(5)}$$

$$r = \sqrt{b_{yx} \cdot b_{xy}} \quad \text{(6)}$$

From (6) we may define the product-moment coefficient of correlation, r, as equal to the geometric mean of the two regression coefficients.

We also note

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = r\frac{s_y}{s_x} \quad \text{(7)}$$

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = r\frac{s_x}{s_y} \quad \text{(8)}$$

and

$$r = \sqrt{\frac{\Sigma xy \cdot \Sigma xy}{\Sigma x^2 \cdot \Sigma y^2}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} \quad \text{(9)}$$

also

$$r = \frac{\Sigma xy}{N s_x s_y}, \text{ the most often given equation for } r \quad \text{(10)}$$

If $r = +1$, we say that the variates are perfectly positively correlated; if $0 < r < 1$, that they are positively correlated; if $r = 0$, that they are uncorrelated; if $O > r > -1$, that they are negatively correlated; and if $r = -1$, that the two variables are perfectly negatively correlated.

To say that two variables are "uncorrelated" does not mean that they are "independent." If the variables are independent they are uncorrelated, but not *vice versa.* A perfect functional relationship may exist between two variables when the product moment r is zero. For example, $y = (\pm x)^2$.

Calculation of the coefficient of correlation in numerical problems will be illustrated first from ungrouped data and then from data grouped into suitable class-intervals.

THE CALCULATION OF THE CORRELATION COEFFICIENT FROM UNGROUPED DATA

An efficient working formula for the calculation of the coefficient of correlation from ungrouped data is

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma(X^2) - (\Sigma X)^2]\,[N\Sigma(Y^2) - (\Sigma Y)^2]}} \quad \ldots\ldots\ldots\ldots (11)$$

An example of the application of (11) is the determination of the coefficient of correlation for a random sample of 30 graduate students between scores on two intelligence tests, Miller's Analogies and the Otis Test (Table 35). The original scores were first reduced: the X's by 60 and the Y's by 50. Substituting the obtained values in formula (11) we have

$$r = \frac{(30)(4565) - (261)(320)}{\sqrt{[(30)(7427) - (261)^2]\,[(30)(4860) - (320)^2]}}$$
$$= .65$$

Provision is made in the last two columns for the check of the primary calculations:

$$\begin{aligned}
\Sigma(x + y + 1) &= \Sigma x + \Sigma y + N \\
&= 261 + 320 + 30 = 611 \\
\Sigma(x + y + 1)^2 &= \Sigma x^2 + \Sigma y^2 + 2\Sigma x + 2\Sigma y + 2\Sigma xy + N \\
&= 7427 + 4860 + 522 + 640 + 9130 + 30 \\
&= 22609
\end{aligned}$$

The test of the significance of the sample coefficient of correlation is given by

$$t = \frac{r\sqrt{N-2}}{\sqrt{1 - r^2}} \quad \ldots\ldots\ldots\ldots (12)$$
$$= \frac{.65\sqrt{28}}{\sqrt{1 - (.65)^2}} = 4.569$$

We enter the table of t with $n = N-2$ (or 28) and find the corresponding value of $P < .001$; therefore, r is significant.

The above method of calculating r is used when the number of observations is small or when a calculating machine is used. The method becomes uneconomical when the number of observations is large and a machine is not available.

CALCULATION OF THE CORRELATION COEFFICIENT FROM GROUPED DATA

We shall illustrate the method of calculating the product-moment coefficient of correlation by obtaining the estimate of the correlation between the height and weight of a representative sample of 690 girls of ages 7.5 to 8.5 years (see Table 36). The variables have been grouped with class intervals of 1 inch for height (X) and

TABLE 35. Calculation of the Product-Moment Coefficient of Correlation from Ungrouped Data: Miller Analogies (X) and Otis Scores (Y)

Ind.	X	Y	x *	y †	x^2	y^2	xy	$(x+y+7)$	$(x+y+7)^2$
1	85	72	25	22	625	484	550	48	2304
2	51	56	−9	6	81	36	−54	−2	4
3	47	60	−13	10	169	100	−130	−2	4
4	80	64	20	14	400	196	280	35	1225
5	67	51	7	1	49	1	7	9	81
6	77	62	17	12	289	144	204	30	900
7	75	62	15	12	225	144	180	28	784
8	76	65	16	15	256	225	240	32	1024
9	53	50	−7	0	49	0	0	−6	36
10	66	54	6	4	36	16	24	11	121
11	56	59	−4	9	16	81	−36	6	36
12	53	57	−7	7	49	49	−49	1	1
13	75	58	15	8	225	64	120	24	576
14	69	47	9	−3	81	9	−27	7	49
15	54	58	−6	8	36	64	−48	3	9
16	72	64	12	14	144	196	168	27	729
17	43	51	−17	1	289	1	−17	−15	225
18	50	54	−10	4	100	16	−40	−5	25
19	99	70	39	20	1521	400	780	60	3600
20	76	69	16	19	256	361	304	36	1296
21	76	56	16	6	256	36	96	23	529
22	69	67	9	17	81	289	153	27	729
23	65	70	5	20	25	400	100	26	676
24	77	64	17	14	289	196	238	32	1024
25	79	68	19	18	361	324	342	38	1444
26	57	54	−3	4	9	16	−12	2	4
27	69	54	9	4	81	16	36	14	196
28	84	64	24	14	576	196	336	39	1521
29	78	70	18	20	324	400	360	39	1521
30	83	70	23	20	529	400	460	44	1936
Total			261	320	7427	4860	4565	611	22609

* $x = X - 60$
† $y = Y - 50$

of 2 pounds for weight (Y). Each pair of measurements was entered as a tally in the appropriate cell of the correlation table. After the total sample of 690 girls had been tabulated in this manner, the tallies in each cell were counted and the number entered. The sums of rows are entered in Column 1; of columns in Row 1; they give the frequency distribution of Y when variation in X is neglected, and of X when variation in Y is neglected.

The working formula for the product-moment coefficient of correlation, which we use here, is:

$$r = \frac{N\Sigma Fxy - (\Sigma Fx)(\Sigma Fy)}{\sqrt{[N\Sigma Fx^2 - (\Sigma Fx)^2]\,[N\Sigma Fy^2 - (\Sigma Fy)^2]}} \quad \ldots\ldots\ldots\ldots (13)$$

TABLE 36. Calculation of the Product Moment Coefficient of Correlation from Grouped Data

X (WEIGHT IN POUNDS)

Y (HEIGHT IN INCHES)

	30.5 32.5	32.5 34.5	34.5 36.5	36.5 38.5	38.5 40.5	40.5 42.5	42.5 44.5	44.5 46.5	46.5 48.5	48.5 50.5	50.5 52.5	52.5 54.5	54.5 56.5	56.5 58.5	58.5 60.5	60.5 62.5	62.5 64.5	F (1)	y (2)	Fy (3)	Fy^2 (4)	Σfx	yΣfx
54 7/8 - 55 7/8							1^{-3}											1	8	8	64	-3	-24
53 7/8 - 54 7/8									$\bar{X}$=47.02									0	7	0	0	0	0
52 7/8 - 53 7/8					1^{-5}										1^{5}			2	6	12	72	0	0
51 7/8 - 52 7/8																		0	5	0	0	0	0
50 7/8 - 51 7/8										1^{0}			1^{3}			1^{6}		3	4	12	48	9	36
49 7/8 - 50 7/8								1^{-2}			1^{1}	1^{2}	3^{9}	3^{12}	2^{10}	1^{6}	1^{7}	13	3	39	117	45	135
48 7/8 - 49 7/8									4^{-4}	1^{0}	4^{4}	6^{12}	10^{30}	3^{12}	3^{15}	3^{18}	2^{14}	36	2	72	144	101	202
47 7/8 - 48 7/8						2^{-8}		1^{-2}	7^{-7}	8^{0}	15^{15}	17^{34}	9^{27}	4^{16}	1^{5}	1^{6}	1^{7}	66	1	66	66	93	93
46 7/8 - 47 7/8							1^{-3}	9^{-18}	17^{-17}	17^{0}	14^{14}	13^{26}	5^{15}	2^{8}	2^{10}			80	0	0	0	35	0
45 7/8 - 46 7/8	$\bar{Y}$=45.63				2^{-10}	4^{-16}	13^{39}	17^{-34}	37^{37}	18^{0}	16^{16}	8^{16}	2^{6}	1^{4}				118	-1	-118	118	(-)94	94
44 7/8 - 45 7/8						19^{-76}	26^{78}	30^{60}	28^{28}	13^{0}	7^{7}	2^{4}	1^{3}	1^{4}				127	-2	-254	508	(-)224	448
43 7/8 - 44 7/8			1^{-7}	3^{18}	8^{-40}	12^{48}	31^{-93}	16^{32}	19^{19}	3^{0}	3^{3}	1^{2}						97	-3	-291	873	(-)252	756
42 7/8 - 43 7/8				5^{30}	15^{75}	13^{52}	13^{39}	8^{-16}	8^{-8}	3^{0}	1^{1}							66	-4	-264	1056	(-)219	876
41 7/8 - 42 7/8			1^{-7}	8^{48}	8^{-40}	8^{32}	2^{-6}	2^{-4}	6^{-6}	2^{0}		2^{4}						39	-5	-195	975	(-)139	695
40 7/8 - 41 7/8		2^{-16}		7^{42}	2^{-10}	4^{-16}	3^{-9}	2^{-4}	1^{-1}									21	-6	-126	756	(-) 98	588
39 7/8 - 40 7/8		1^{-8}	3^{-21}	1^{-6}	2^{10}	2^{-8}		1^{-2}	2^{-2}									12	-7	-84	588	(-) 57	399
38 7/8 - 39 7/8	1^{-9}			1^{-6}	1^{-5}						1^{1}							4	-8	-32	256	(-)19	152
37 7/8 - 38 7/8		2^{-16}																2	-9	-18	162	(-)16	144
36 7/8 - 37 7/8		1^{-8}			1^{-5}	1^{-4}												3	-10	-30	300	(-)17	170
(1) F	1	6	5	25	40	65	90	87	129	66	62	50	31	14	9	6	4	690 (N)		-1203 (ΣFy)	6103 (ΣFy^2)	-855	4764 (ΣFxy)
(2) x	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7						
(3) Fx	-9	-48	-35	-150	-200	-260	-270	-174	-129	0	62	100	93	56	45	36	28	-855 (ΣFx)					
(4) Fx^2	81	384	245	900	1000	1040	810	348	129	0	62	200	279	224	225	216	196	6339 (ΣFx^2)					

The quantities needed for substitution in (13) are calculated by using the columns to the right of and the rows at the bottom of the correlation table in the following manner:

> F (Column 1) denotes the frequency for each row
> y (Column 2) denotes the computation variable for Y
> Fy is the product of Columns 1 and 2
> Fy^2 is the product of Columns 2 and 3
>
> Σfx for each row is obtained by adding all the values in the upper right-hand corner of each cell for each row. The value for any given cell is obtained by multiplying the cell frequency f by the value of the computation variable x for the column in which the cell lies. Thus in the first vertical column of the correlation table multiply the cell frequency by -9; in the second column multiply the cell frequency by -8; and so forth.
>
> By way of explanation of the entries under Σfx, note the sixth entry, 45. This comes from
> $1\text{x}-2+1\text{x}1+1\text{x}2+3\text{x}3+3\text{x}4+2\text{x}5+1\text{x}6+1\text{x}7$*.
>
> $y\Sigma fx$ is the product of columns 2 and 5.
>
> From the rows at the bottom of the table we have
> F (Row 1) denotes the frequency for each column
> x (Row 2) denotes the computation variable for X
> Fx is the product of Rows 1 and 2
> Fx^2 is the product of Rows 2 and 3.

Summing the values in the appropriate columns and rows we have for our problem

$$N = 690 \qquad \Sigma Fx^2 = 6339$$
$$\Sigma Fxy = 4764 \qquad (\Sigma Fx)^2 = (-855)^2$$
$$(\Sigma Fx) = -855 \qquad \Sigma Fy^2 = 6103$$
$$(\Sigma Fy) = -1203 \qquad (\Sigma Fy)^2 = (-1203)^2$$

Substituting these values in Eq. (13) we obtain the value of r_{xy}:

$$r_{xy} = \frac{690 \times 4764 - (-855)(-1203)}{\sqrt{[690 \times 6339 - (-855)^2]\,[690 \times 6103 - (-1203)^2]}}$$
$$= .7118$$

To test the significance of $r = .71$, we obtain

* After one gains experience it is less necessary to insert the upper numbers in the cells, or one can use a strip containing the computation variable for x and moving it down a row at a time secure the sum of the products.

$$t = \frac{.71\sqrt{688}}{\sqrt{1 - (.71)^2}}$$
$$= 26.4$$

With $d.f. = 688$, $P < .001$, $\therefore$ r is highly significant.

PRACTICAL EXAMPLE OF CORRELATION THEORY

We shall make use of the problem just given to illustrate the calculation of the coefficient of correlation from grouped data (Table 36) for further illustration of the theory discussed in the first section.

In Table 36, we have drawn two regression lines, one of which is the regression line of Y on X, the other the regression line of X on Y.

The equations of the two lines are

$$(Y - \bar{Y}) = r\frac{\sigma_y}{\sigma_x}(X - \bar{X}) \quad \ldots\ldots (14)$$

$$(X - \bar{X}) = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y}) \quad \ldots\ldots (15)$$

where X and Y are values given by the line, $\bar{X}$ and $\bar{Y}$ are the two means, σ_x and σ_y are the two standard deviations, and r is the correlation coefficient.

Equation (14) may be used to predict the mean value of Y for any value of X, and is the line more nearly parallel to the X-axis. Equation (15) represents the other line; both pass through the point $(\bar{X}, \bar{Y})$.

It will be noted that equations (14) and (15) are not algebraic; one cannot be derived from the other. Since the relationship between the equations is not a reciprocal one the same equation cannot be used for predicting Y from X and X from Y. This is because the relations given in these equations represent average and not individual results.

The two regression lines were fitted by the method of least squares where the slopes, (b_{yx}, b_{xy}), and the means $(\bar{X}, \bar{Y})$ were the least square estimates.

The calculation of the slope by the method of least squares has been previously described. We shall use this method rather than calculate $r\frac{\sigma_y}{\sigma_x}$ and $r\frac{\sigma_x}{\sigma_y}$ as given in Equations (14) and (15).

For our problem:

$$b_{yx} = \frac{\Sigma Fxy - \dfrac{(\Sigma Fx)(\Sigma Fy)}{N}}{\Sigma Fx^2 - \dfrac{(\Sigma Fx)^2}{N}} \cdot \frac{(h)(k)}{(h)^2}$$

$$= \frac{4764 - \dfrac{(-855)(-1203)}{690}}{6339 - \dfrac{(-855)^2}{690}} \cdot \frac{(2)(1)}{(2)^2}$$

$$= .310$$

$$b_{xy} = \frac{4764 - \dfrac{(-855)(-1203)}{690}}{6103 - \dfrac{(-1203)^2}{690}} \cdot \frac{(1)(2)}{(1)^2}$$

$$= 1.634$$

$$s_x = h\sqrt{\frac{\Sigma Fx^2}{N-1} - \frac{(\Sigma Fx)^2}{N(N-1)}} = 2\sqrt{\frac{6339}{689} - \frac{(-855)^2}{689(690)}}$$

$$= 5.536$$

$$s_y = k\sqrt{\frac{\Sigma Fy^2}{N-1} - \frac{(\Sigma Fy)^2}{N(N-1)}} = 1\sqrt{\frac{6103}{689} - \frac{(-1203)^2}{689(690)}}$$

$$= 2.411$$

$$\bar{X} = 49.50 + \frac{2(-855)}{690} = 49.50 - 2.48\ (=)\ 47.02$$

$$\bar{Y} = 47.375 + \frac{(-1203)}{690} = 47.375 - 1.743\ (=)\ 45.63$$

From Equation (6) we find

$$r = \sqrt{b_{xy} \cdot b_{yx}}$$
$$= \sqrt{.310 \cdot .634}$$
$$= .7118$$

The regression equations for Y on X and for X on Y in raw score form are given by:

$$Y_E = \bar{Y} + b_{yx}(X - \bar{X})$$
$$= 45.63 + .310\,(X - 47.02)$$
$$= 30.87 + .310X$$
$$X_E = \bar{X} + b_{xy}(Y - \bar{Y})$$
$$= 47.02 + 1.634\,(Y - 45.63)$$
$$= -27.54 + 1.634Y$$

We may also fit the regression line to the points determined by the means of the Y-distributions against the centers of the corresponding X-intervals. Similarly, the other regression line may be

fitted to the means of the X-distributions against the centers of the Y-intervals.

The calculations for the slope of the regression line for Y and X are given in Table 37.

TABLE 37. Fitting the Regression Line for Estimating Y from X Using the Means of the Columns of the Y Variable and the Mid-points of the X-Variable

X_{mp}	$\bar{Y}$
33.5	40.88 *
37.5	42.34
39.5	43.28
41.5	44.05
43.5	44.82
45.5	46.36
47.5	45.69
49.5	46.45
51.5	46.97
53.5	47.52
55.5	48.60
57.5	48.52
61.5	50.00 *

$$N = 13$$

$$\Sigma\bar{Y} = 595.48 \qquad \bar{\bar{Y}} = 45.81$$

$$\Sigma X_{mp} = 617.5 \qquad \bar{X}_{mp} = 47.50$$

$$\Sigma X_{mp}\bar{Y} = 28546.940$$

$$\Sigma X_{mp}^2 = 30163.25$$

$$b_{xy} = \frac{N\Sigma X_{mp}\bar{Y} - (\Sigma X_{mp})(\Sigma\bar{Y})}{N\Sigma X_{mp}^2 - (\Sigma X_{mp})^2}$$

$$b_{xy} = \frac{3401.320}{10816.00} = .3145$$

$$\bar{Y}_E = \bar{\bar{Y}} + b_{yx}(X_{mp} - \bar{X}_{mp})$$

$$\bar{Y}_E = 30.87 + .3145\, X_{mp}$$

* Weighted means.

The corresponding calculations for X on Y are given in Table 38.

The two fitted regression lines are observed in Figure 6. However, they are fitted to the means of raw scores based on different scales of measurement.

We noted in the theoretical discussion (Fig. 6) that it was necessary to render the slopes of the regression lines independent of the units of measurement used for X and Y. This is done by converting the original measures to standard measure and then fitting the regression lines to the means of the standard scores of the respective rows and columns. This has been done. The calculations for b_{yx} established on the column means of standard scores are given in Table 39. Similarly, the calculations for b_{xy} are found in Table 40.

In Figure 7, we have fitted the regressions lines, one to the means of column, the other to the means of rows where the mid-points and means are expressed in standard measure form.

TABLE 38. Fitting the Regression Line for Estimating X from Y Using the Means of the Rows of the X Variable and the Mid-points of the Y-Intervals

$\bar{X}$	Y_{mp}
61.50 *	53.375
56.42	50.375
55.11	49.375
52.32	48.375
50.38	47.375
47.91	46.375
45.97	45.375
44.30	44.375
42.86	43.375
42.37	42.375
40.17	41.375
40.00	40.375
35.50 *	38.375

$N = 13$

$\Sigma\bar{X} = 614.81 \qquad \bar{\bar{X}} = 47.29$

$\Sigma Y_{mp} = 590.875 \qquad \bar{Y}_{mp} = 45.45$

$\Sigma\bar{X}Y_{mp} = 28330.86375$

$\Sigma\bar{X}^2 = 29829.2607$

$\Sigma Y_{mp}^2 = 27079.328125$

$$b_{xy} = \frac{N\Sigma\bar{X}Y_{mp} - (\Sigma\bar{X})(\Sigma Y_{mp})}{N\Sigma Y^2_{mp} - (\Sigma Y_{mp})^2}$$

$$b_{xy} = \frac{5025.37}{2898.00} = 1.734$$

$$\bar{X}_E = 47.29 + 1.734\,(Y_{mp} - 45.45)$$
$$= -31.52 + 1.734\,(Y_{mp})$$

* Weighted means.

TABLE 39. Fitting the Regression Line for Estimating Y from X Using the Means of Standard Scores of the Y Variable in the Columns and the Standard Scores of the Mid-points of the X-Variable

X_{mp}	$\bar{Y}$
+2.615	1.800 *
+1.893	1.197
+1.531	1.231
+1.170	.781
+.809	.556
+.447	.340
+.086	.026
−.274	−.143
−.635	−.335
−.997	−.654
−1.358	−.976
−1.719	−1.366
−2.442	−1.800 *

$N = 13$

$\Sigma X_{mp} = +1.126$

$\Sigma X_{mp}^2 = 27.231020$

$\Sigma\bar{Y} = +.657$

$\Sigma X_{mp}Y = 19.348479$

$$b_{yx} = \frac{N\Sigma X_{mp}\bar{Y} - (\Sigma X_{mp})(\Sigma\bar{Y})}{N\Sigma X_{mp}^2 - (\Sigma X_{mp})^2}$$

$$b_{yx} = \frac{250.790}{352.735} = .7110$$

$$\bar{Y}_E = \bar{\bar{Y}} + .7110\,(X_{mp} - \bar{X}_{mp})$$
$$= .051 + .7110\,(X_{mp} - .087)$$
$$= -.011 + .7110\,X_{mp}$$

* Weighted means.

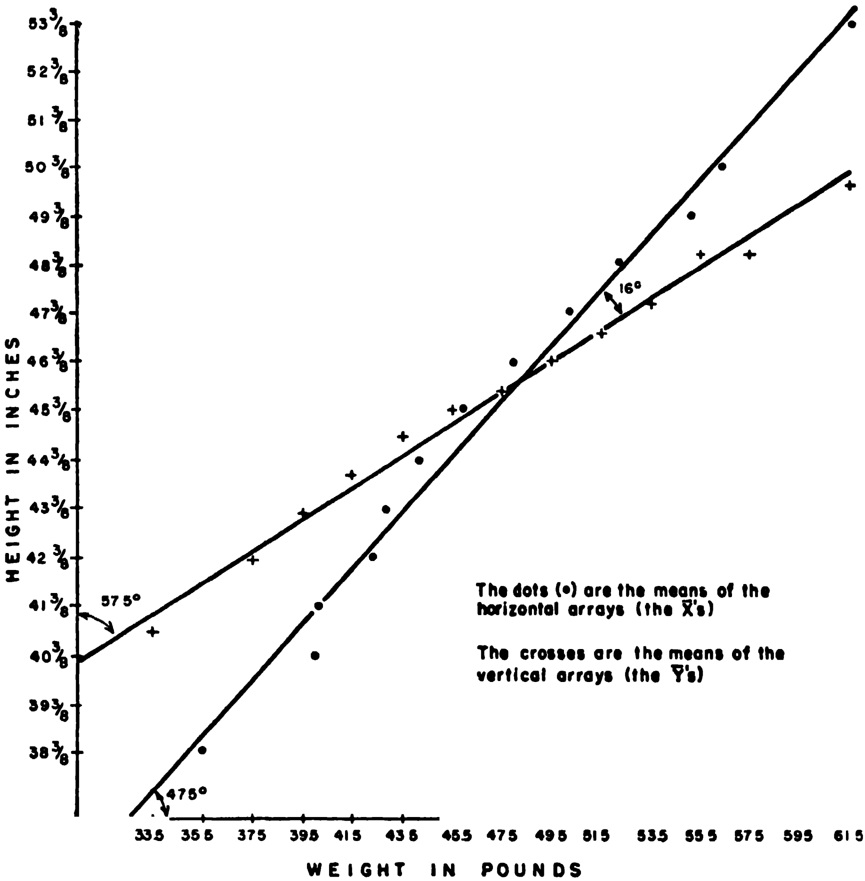


FIGURE 6. Regression line fitted to means of the raw scores in rows and columns.

The equations for these regression lines fitted to standard scores are:

$Y_E = -.011 + .7110\, X$(midpoint value of X in standard measures)
$X_E = .070 + .7086\, Y$(midpoint value of Y in standard measures)

It is now observed that the tangent of the angle between each regression line and the axis of the corresponding independent variable each becomes equal to the coefficient of correlation. Thus

$$\text{Tan}\,\alpha = \text{Tan}\,\beta = \text{Tan}\,35°30' \qquad (16)$$
$$= .71$$

It is also worth noting that the cosine of the angle between the two regression lines has the following functional relationship with the coefficient of correlation.

$$\text{Cos}\,\gamma = \frac{2r}{r^2+1};\ \text{for } r = .71 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (17)$$

$$\text{Cos}\,\gamma = \frac{1.42}{1.5041}\ (=)\ .9441$$

$$\gamma = 19°15'$$

Which is very close to the observed angle γ in Figure 7.[1] The cosine of the angle between the two regression lines fitted to the means

TABLE 40. Fitting the Regression Line for Estimating X from Y Using the Means of Standard Scores of the X Variable in the Rows and the Standard Scores of the Mid-points of Y-Interval

Y_{mp}	$\bar{X}$
3.212	2.150 *
1.968	1.698
1.553	1.460
1.138	.967
.723	.576
.309	.159
−.105	−.198
−.520	−.490
−.935	−.750
−1.350	−.839
−1.764	−1.237
−2.179	−1.267
−3.009	−2.000 *

$$N = 13$$

$$\Sigma \bar{X} Y_{mp} = 27.151220$$

$$\Sigma Y_{mp} = -.949$$

$$\Sigma \bar{X} = +.229$$

$$\Sigma Y_{mp}^2 = 38.406999$$

$$b_{xy} = \frac{N\Sigma \bar{X} Y_{mp} - (\Sigma \bar{X})(\Sigma Y_{mp})}{N\Sigma Y_{mp}^2 - (\Sigma Y_{mp})^2}$$

$$= \frac{353.183}{498.390} = .7086$$

$$\begin{aligned} X_E &= \bar{X} + .7086\,(Y_{mp} - \bar{Y}_{mp}) \\ &= .018 + .7086\,(Y_{mp} + .073) \\ &= .070 + .7086\,Y_{mp} \end{aligned}$$

* Weighted means.

of standard scores is a convenient function to use since it ranges in value from −1 to +1 as does the coefficient of correlation.[2]

THE USES AND ABUSES OF CORRELATION

The validity of the test of significance of the correlation coefficient and of its estimation (either point or interval) is contingent

[1] The reason for the slight discrepancy between the observed angle $\gamma = 19°$ and the theoretical value 19°15′, and for the slight displacement of the intersection of the regression lines and the intersection of the X, Y axes is that the standard scores were calculated from the grouped data rather than from the original measures.

[2] The coefficient of correlation between two variables expressed as deviations from their respective means is the cosine of the angle between their vectors in n- dimensional space.

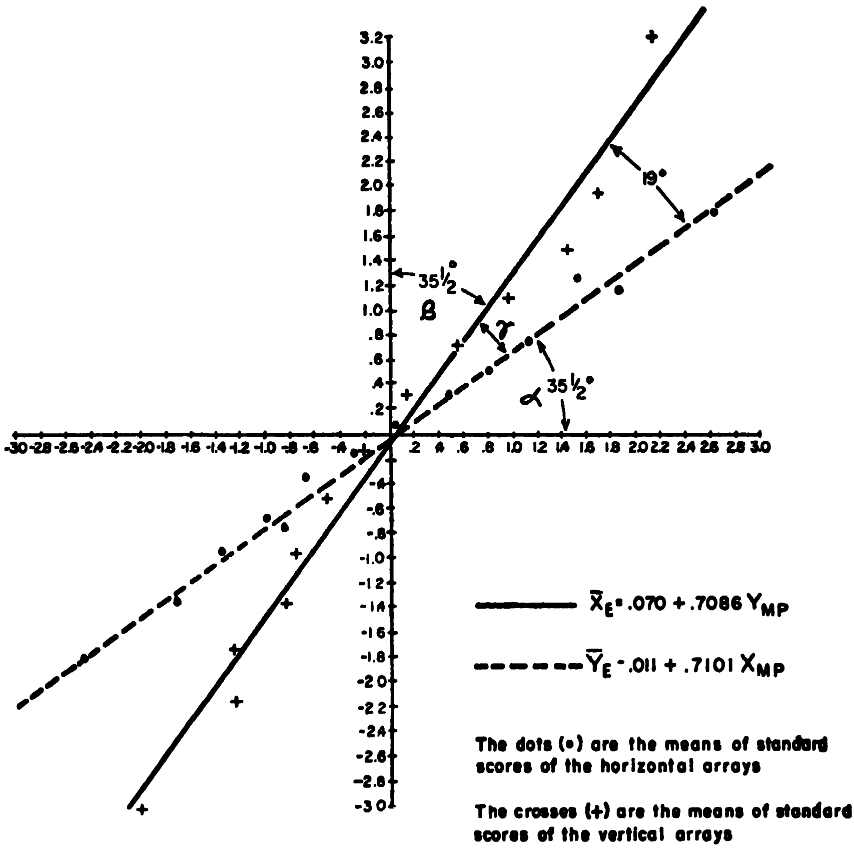


FIGURE 7. Regression lines fitted to means of the standard scores.

upon the selection of the sample by the method that will best ensure its randomness and representativeness.

The critical research worker is continuously attentive to factors that might lead to biased or spurious estimates. Some of these factors will be briefly discussed.

1) **Selection of the sample may be made in such a way as to exclude values at the extremes or to exclude middle values and include the extremes.** The effect in the former case is to lower the correlation coefficient; in the latter case, to raise the coefficient.

2) **The data with which the investigator deals are subject to errors of observation and measurement.** Errors of this type may or may not be correlated with the value being measured. If uncorrelated they tend to fall equally above and below the true value throughout the range of the variable. The effect of this type of error is to lower the value of the correlation coefficient. Various

methods are available for determining the reliability of tests.[1] If an estimate of the correlation between perfectly reliable measures of the variable is desired, formulas are available to correct for attenuation.

When the errors are correlated with the variable, the tendency is usually to make the observed value fall above the true value in the upper part of the range, and below in the lower part; or vice versa. The effect of this type of error, which is called systematic error, is to obscure the true relationship and give biased estimates of the correlation coefficient. The tendency is to make the estimates either an over- or under-estimate, depending on the interrelations between the variables and the errors.[2]

3) **Two characteristics can be correlated because they both are affected by a third group of factors. Sometimes they just happen to be correlated.** Means are available for isolating and measuring more exactly the net correlation between two variables affected by other common variables. The technique is that of partial correlation.

4) **The efficient use of the product-moment correlation coefficient depends upon the fulfillment of certain assumptions:** the linearity of regression; that is, the regression line of each variate on the other is straight, and the variance of X is the same for all arrays, and so for Y. When the regression is nonlinear, it is necessary to use some other type of equation that will give a good fit if the functional relation is to be used for prediction purposes. The correlation ratio is sometimes used as a descriptive statistic.

The effect on the correlation coefficient, when the assumption of linearity is not fulfilled, is to underestimate its value. Very little knowledge about the sampling distribution of the correlation coefficient is available except in the case of the normal parent or population. Some empirical results show that a considerable deformation in the normal correlation surface may be tolerable. The condition of homoscedasticity is characteristic of the bivariate normal distribution. This means that the variance of X-arrays is the same for all arrays; likewise for Y.

In correlation procedure (likewise for regression), one of the fundamental assumptions involves the independence and randomness of the observations. Certain types of data are not independent of time and space. Most time series should be examined critically

[1] Palmer O. Johnson, *Statistical Methods in Research* (New York: Prentice-Hall, 1949), pp. 125–138.

[2] Mordecai Ezekiel, *Methods of Correlation Analysis* (New York: John Wiley and Sons, 1930), Chapter 19.

for the fulfillment of conditions of randomness and independence of observations, e.g., the weekly observations on the mental or physical growth of a child would not be independent of each other. There are also various types of data that may have serial correlation because of position in space.[1]

5) **Erroneous conclusions are sometimes drawn from data of quantities that change in time; in some types of data it is necessary to recognize "nonsense correlations."** For example, a coefficient of correlation of .86 was found between variations in the death rate of the state of Hyderabad from 1911 to 1919 and variations in the memberships of the International Association of Machinists from 1912 to 1920.

6) **Caution is necessary for the correlation analysis of measures which may be in the form of indices.** If all the separate factors and their interrelationships are not allowed for, spurious coefficients of correlation may be obtained. Examples are: estimating the relationship between two sets of measurements obtained by correlating I.Q. scores, correlating I.Q. scores with chronological age, or correlating educational quotients with accomplishment quotients.[2] Chronological age is the common factor responsible for the spurious correlation.

There are other difficulties involved in the interpretation of coefficients of correlation. The more common of these are:

1) Failure to take into account the unevenness of the correlation scale. Note, for example, the data in Table 41, in which the ratio of the residual to the total mean variance and the ratio of the corresponding standard deviations are given for specified values of r. It will be noted that the scale of r is very uneven. For example, a correlation coefficient of 0.6 reduced the residual standard deviation to 0.8 of the total while even for a correlation as high as 0.90, the residual deviations have a standard deviation of 0.436 of the total. For a correlation of 0.866 the standard deviation is .50 of the total. Failure to take into account this unevenness of the correlation scale is often a source of error in interpreting the relative importance of coefficients of different sizes.

2) Interpreting the coefficient of correlation as a percentage. The square of r, or r^2, under certain conditions may be regarded as representing the proportion of the variance in one variable accounted for by that of the other. This is based upon the additive property of

[1] Lila F. Knudsen, "Interdependence in a series," *Journal of American Statistical Association,* 1940, 35:507–517.

[2] Robert W. B. Jackson, "Some pitfalls in statistical analysis of data expressed in form of I.Q. scores," *Journal of Educational Psychology,* 1940, 31:677–685.

the variance, that is, the variance of the residuals plus the variance of one variable (say, the independent variable X) equals the variance of the other variable (say, the dependent variable Y).

3) *Interpreting the coefficient as cause and effect.* A correlation may result directly from the operation of a cause but its existence does not prove a causal relationship. The deduction of causality might be drawn from a very high correlation. For example, an r of .9919 has been reported between temperature in degrees Fahrenheit and the number of chirps of crickets per minute corresponding to the temperature reading. Such a deduction, even when dealing with quantities capable of close control, is often erroneous. It is especially hazardous when there are variations which are uncontrolled, and the relationship is not exact. Only when all of the possible factors that might contribute to variation have been taken into account can a causal relationship be reasonably inferred. When causation is under investigation, it is good practice to decide first on other grounds that a causal relationship is probable. One should then analyze critically other possible factors before presenting the correlation coefficient as evidence.

TABLE 41. Values of the Residual Deviations for Different Values of the Correlation Coefficient

Correlation Coefficient	*Ratio of Variances Residual/Total*	*Ratio of Standard Deviations Residual/Total*
1.00	0.00	0.000
0.90	0.19	0.436
0.80	0.36	0.600
0.70	0.51	0.714
0.60	0.64	0.800
0.40	0.84	0.917
0.20	0.96	0.980
0.00	1.00	1.000

In considering the possible effect of other factors as influencing the value of the correlation coefficient obtained between two variables, the partial correlation technique can be used. For example, a physiologist obtained a negative correlation between the quantity of calcium in the bones of a series of persons and the number of living aunts. Obviously there is no direct causal relation. Calcium content and number of aunts are related to age. Bones of older people have more calcium than do bones of younger persons; also older people are likely to have a smaller number of surviving aunts.

If all individuals in the sample had been of the same age, the correlation would likely have been zero. The same result would have been obtained with the application of partial correlation by which it is possible (for example, where the correlation between each pair of three variates has been found) to obtain the direct correlation between any two by holding the third variate constant. Likewise, two or more variates may be eliminated in succession. Caution is required in the use of the partial correlation technique. The logic underlying the technique is that the control of the variate or variates eliminated is equivalent to experimental controls. The validity of this relation should be examined before applications are made.

4) *Problems arise where the research worker needs to determine the relation of a part to the whole.* For example, one may determine the relation between heart weight and body weight. This problem is most efficiently handled by analysis of covariance techniques.[1]

5) *Use often is made of such statistical constants as the coefficient of alienation,* $k = \sqrt{1 - r^2}$, and the index of forecasting efficiency, $E = 1 - k = 1 - \sqrt{1 - r^2}$, when interpreting the correlation coefficient for purposes of prediction. Actually these statistics are measures of the residual deviation about the regression lines and as measures of "efficiency" they yield a misleading result. Even a correlation coefficient as low as 0.30 can be of considerable value in predicting success or failure. Assuming that one would wish to select on the basis of the criterion measure by setting up a *failure ratio of 30,* it would be found that the efficiency of predictions of failures would range from 62.8 per cent in the lower 10 per cent of the distribution of the variable used as a predictor to 92.4 in the upper 10 per cent of the distribution. By a failure ratio is meant the percentage of the whole group, classified as unsuccessful or failures on whatever criterion may be used. Tables which give the prediction efficiencies by deciles for various degrees of relationships have been prepared.[2]

The measure of relationship, which we have discussed, is the Pearson product-moment coefficient of correlation. It is in general the best measure of mutual linear relationship between two varia-

[1] R. A. Fisher, "The analysis of covariance method for the relation between a part and the whole," *Biometrics,* 1917, 3:65–68.

[2] Robert W. B. Jackson and Alexander J. Phillips, *Prediction Efficiencies by Deciles for Various Degrees of Relationship* (Toronto: Department of Educational Research, University of Toronto); and Reigh H. Bittner and Carlton E. Wilder, "Expectancy tables: a method of interpreting correlation coefficients," *Journal of Experimental Education,* 1946, 14:245–252.

bles. The utility of this method of correlation was extended in its application to groups of more than two variates. The product-moment coefficient of correlation has been a useful research tool, particularly because we have a knowledge of its sampling distribution; appropriate tests of significance and methods of estimation are available.

A number of other methods of measuring association have been developed chiefly for situations to which the product-moment correlation coefficient is not appropriate. Among these methods are the coefficient of contingency, tetrachoric, biserial, and the like. These statistical methods do not share the advantages of the correlation coefficient as an instrument of science. Practically nothing is known about their sampling distributions. Hence exact tests of significance are not available. They should be used as sparingly as possible and rarely, if ever, as ends in themselves. There are certain psychological problems in testing and elsewhere, where, regardless of their unreliability, they seem to be the only tools available. On the other hand they have often been used where other methods of analysis are much more appropriate and efficient. This latter statement also applies to the product-moment correlation coefficient. The tradition of working out coefficients of correlation frequently has led workers to calculate them when the data could have been more appropriately and efficiently analyzed by other methods. It is not possible here to discuss all the short-cut methods of correlation that are available. The student is referred to *Fundamentals of Statistics*.[1]

We shall conclude our discussion of correlation analysis by summarizing briefly the role of the correlation coefficient in research.

THE CORRELATION COEFFICIENT IN RESEARCH

We should like to point out that regression has proved to be a more generally useful tool than correlation. We have treated this problem in Chapter IX.

It has been found that most, though not all, correlation problems arising in practice can be dealt with more appropriately by regression methods. The correlation coefficient provides no new information, once the regression coefficient and the two components of the sum of squares of the dependent variable, say Y, are known. Since these quantities are almost always needed, r becomes superfluous. Moreover, in many experimental data, the values of the independent variable, say X, are selected. The regression coefficient does not change with the distribution of X, insofar as the values of

[1] Truman L. Kelley, *Fundamentals of Statistics* (Cambridge: Harvard University Press, 1947).

X are selected within the range where regression is linear. On the contrary, the coefficient of correlation changes with the distribution of X-values. Also of great theoretical and practical importance is the fact that regression methods are based only on the assumption that deviations from the regression function be normal; whereas in correlation analysis, it is required that the variate and certain usually observable and known parameters for the given sample all be jointly normally distributed.

On the other hand, correlation, when conditions permit (i.e., where the data are accurate, homogeneous, representative, or unselected) permits wide generalizations. Some major generalizations are based on correlation analysis in fields where experimentation would be difficult, if not impossible. Thus R. A. Fisher showed in 1918 that the biometrical correlations observed between relatives were those to be expected on the Mendelian theory. Negative correlation between intelligence and family size was strongly confirmed in the Scottish survey (1947). This correlation persisted both among children with very young and with relatively old mothers, and also with separate occupational groups. Thus, there was the same reduction of 1½ points in score (roughly equivalent to 1½ I.Q. points) for each additional child in the professional group as in the whole sample. These findings illustrate the significant contribution of correlation analysis in situations over which we cannot usually exercise experimental controls. That is, the analysis of data where we can observe the occurrences of various possible causes that may contribute to the explanation of a phenomenon but cannot control the factors. Correlational analysis, where experimentation is possible, can also serve to identify factors that are worthy of experimentation.

Perhaps the best-known uses of the correlation coefficient are in connection with the theory and practice of measurement. Reliability coefficients have been previously referred to, as have validity coefficients. Considerable study has been given to the factors affecting the values of these coefficients. Other direct uses of the product moment correlation coefficient are in the determination of the relative intensity of relations among variables, of the relation of grades in one subject to grades in another, of scores on one mental test to scores on another, and of mental and physical characteristics.[1]

[1] The following articles and books have dealt with such correlations:

American Educational Research Association, "Psychological Tests and Their Uses," *Review of Educational Research*, 1947, 17.

Paul Blommers and E. F. Lindquist, "Rate of comprehension of reading," *Journal of Educational Psychology*, 35: 449–73 (November, 1944).

Harry S. Dyer, "Validity of certain objective techniques for measuring the ability

An important development in psychological research which originated from the study of correlation coefficients was the inquiry of Spearman as to whether all intellectual abilities could be resolved into a single general ability and a number of separate specific abilities. Subsequent investigations with psychological data have not found the single factor hypothesis adequate for mental testing. This circumstance has led to the development of more complex statistical techniques designed to differentiate more than two factors in an analysis of the tables of the correlation coefficients between mental tests. L. L. Thurstone, T. L. Kelley, C. Burt, H. Hotelling, and others have devised methods for carrying out this multiple-factor analysis. The factorial method most commonly employed has been Thurstone's centroid technique. His approach is to consider a set of more than one common factor. The assumption is that if all the common factors are held constant, then the partial correlations between the original measures vanish. A limitation of the multiple-factor approach is the fact that the correlations between an observed variable and a common factor are not uniquely determinable. There exists infinitely many sets of factors which will satisfy the conditions for vanishing partial correlations. Thurstone attempts to remove the conditions of indeterminacy by imposing a certain criterion, i.e., that certain of the correlations vanish in order to obtain the so-called "simple structure"—: a set of orthogonal common factors, each accounting for 10 to 20 per cent of the test variances. By orthogonal is meant uncorrelated factors. This method does not insist on orthogonality after rotation. Nor does it permit of a general factor.

Generally speaking, the method of factor analysis has been employed as a tool for discovering hypotheses rather than for testing well-defined hypotheses concerning the structure of the variables analyzed. The quantities to be analyzed are subject to fluctuations and consequently to sampling errors, as is true for any function of these quantities. There are essentially two sampling problems, the sample of persons who took the test, and the sampling of a popula-

---

to translate German into English," *Journal of Educational Psychology*, 1946, 37:171–178.

Harold Gulliksen, *Theory of Mental Tests* (New York: John Wiley and Sons, 1950).

M. P. Honzik and J. W. Allen, "The stability of mental test performance between two and eighteen years," *The Journal of Experimental Education*, *17:* 309–15 (December, 1948).

Frank Sandon, "Selection by a nearly perfect examination," *Annals of Eugenics*, 7: 67–85 (June, 1936).

William W. Turnbull, "The relationship between verbal factor scores and other variables," pp. 54–55 in *Proceedings of the 1948 Invitational Conference on Testing Problems* (Princeton, N. J.: Educational Testing Service, 1948).

tion of tests. The sampling problems have been little explored up to this time. Under certain conditions there is a method now available for testing how many common factors can be considered significant.[1] In some investigations more factors have been extracted than the size of the data would justify.

Lack of appropriate tests of significance is a severe limitation on the method of factor analysis as an instrument of science. There have also been applications of the methods in which certain short-approximate methods of measuring correlation have been used as substitutes for the product-moment correlation coefficient which for all theory so far developed has been assumed. The conditions, if any, under which such substitutes may be validly used have not been formulated.

It may be said, however, that correlation and its factorial developments have contributed to solving problems dealing with theories regarding mental organization and with individual differences.[2]

The entire field of multivariate statistical theory is undergoing rapid development. This development has made available new exact tests of statistical hypotheses over a wide range of fields in which multiple measurements are involved. These developments have been largely along the lines of multiple regression discussed in the previous chapter. During the last thirty years there has been a shift in emphasis from correlation coefficients to regression coefficients, as well as from bivariate to multivariate analysis. Corresponding changes have taken place resulting in the modern design of experiments in which the multifactor experiment is rapidly replacing the traditional single-factor experiment.

SUMMARY

Correlation is an extension of regression theory. The theoretical basis for this extension is given and the theoretical steps are illustrated in the solution of a practical problem. Correlation analysis is applied when two or more characteristics are measured on each individual and when it is necessary to take account of the relationship between the variables. Efficient methods of calculating and

[1] Paul G. Hoel, "A significance test for minimum rank in factor analysis," *Psychometrika*, 1939, 4:245–253.

[2] William E. Hall, "An analytical approach to the study of reading skills," *The Journal of Educational Psychology*, 1945, 37:429–442. Karl J. Holzinger and Harry H. Harman, *Factor Analysis* (Chicago: The University of Chicago Press, 1941); L. L. Thurstone, *Multiple Factor Analysis* (Chicago: The University of Chicago Press, 1947).

checking the product-moment correlation coefficient are illustrated for both ungrouped and grouped data. Appropriate tests of significance are made. It is good practice to plot the observational values in a scatter diagram to obtain preliminary information about the degree of relationship between two variables. Grouping the data into class intervals requires less time than plotting the individual data on graph paper. Preliminary information about the fulfillment of the assumptions underlying the product-moment correlation coefficient also becomes available. The assumptions of linearity of regressions and the equality of variances of the different arrays need to be tested.

The correlation coefficient is useful as a shorthand description of the intensity of correlation, but experience and much caution are required to appreciate it. The strength of correlation and its statistical significance are quite different matters. Many factors may affect the magnitude of the correlation coefficient and its interpretation. This chapter enumerates and discusses many practical situations that the sophisticated research worker should be aware of if the correlation coefficient is to be appropriately used.

CHAPTER XI

Complex Developmental Studies

Up to this point we have been concerned with certain types of relationship that frequently become the object of investigation, namely: (1) interrelationships among products, processes and conditions; (2) the relationship between present and future status; and (3) the relationships between samples and population. The methods and techniques discussed in these chapters are varied and complex but indispensable in making many types of investigations and in establishing a foundation for others. The purpose of this chapter is to discuss the interrelationships among the many methods and techniques of appraisal and research discussed in these and earlier chapters and to indicate how they may be combined in meaningful studies of the problems of education.

Of the many methods and techniques of research and appraisal discussed some are more appropriate for some purposes than others. There are, for example, four points of view from which status may be appraised: (1) of educational needs and purposes; (2) of the persons involved; (3) of the principles of action that one holds to be true; and (4) of setting, situation, or foundations. Each of these points of view has been associated with a characteristic method of research. Survey techniques, for example, have been quite generally related to the purposes of education, whether explicitly stated or not. The clinical and diagnostic techniques, although permitting wide applications, have been closely related to persons. The experimental and statistical techniques have been widely applied to a variety of problems, but more particularly to those of interrelationship. The historical and sociological techniques have been applied to settings and social founda-

tions. The truths established by these several disciplines, methods, and techniques of research are, however, frequently true only within circumscribed limits. Many activities are appropriate only to the purposes and conditions that give rise to them; many generalizations true for groups are not true for all the members of the groups. Many inferences too are true only for the peoples, countries, and cultures that give rise to them and only under the particular conditions present. Many inconsistencies in the findings of research workers might be resolved through the use of more inclusive methods of research and appraisal.

Thesis advisers have long advised graduate students to delimit the problem to be studied and sharply narrow the field of search. Possibly we might better give them the diametrically opposite admonition. More accurately they need to do both; i.e., to be much more inclusive in their thinking and in getting a background for their research, later intensifying the search in particular directions. Our research and appraisal activities should be better oriented; a tentative generalization derived from one discipline should be checked against that of others. Specialists in various fields need to work together and appreciate the approximate character of their generalizations. The purposes for which research workers conduct research and appraisal activities are various but interrelated. Hence the necessity for complex studies and well-planned programs of evaluation. The purpose of this chapter is to discuss some of the more inclusive attempts at research and appraisal and outline certain generalizations that may provide a summary of what has preceded.

CONSIDERATIONS FOR IMPROVEMENT

Numerous suggestions have been made throughout this volume for the improvement of appraisal activities. Among the more important of these are the following:

1) Educational needs and purposes should be carefully defined; they can be defined in terms of qualities of the person, behavior, or controls over behavior. Appraisals arising from different statements should be consistent.
2) Data-gathering devices must be carefully validated for the specific purposes for which they are used in particular situations.
3) The methods of recording data should be objectified. Much of the symbolism now employed is ambiguous.
4) Facts are relative; they are always in context; to separate them

from their context may lead to error; the limits within which they may be true should be carefully stated.

5) A clear distinction should be made between descriptive and inferential research; many reliability formulas are based upon sampling and not applicable except where the conditions of probability sampling are observed.
6) Every research involves assumptions; these assumptions must be carefully stated, and their implications checked in the development of investigational designs, in the collection and analysis of data, and in the statement of generalizations.
7) Decisions must be made with reference to the methods and techniques to be employed. Where there are a number of equally appropriate means of analyzing data, the results secured from different approaches should check.
8) Generalizations need to be carefully restricted,—that is restricted to stated purposes, assumptions, courses of action, persons, and conditions.
9) Only where there are samples properly drawn from well-defined populations may the results observed for the samples be generalized to include the population.
10) The implications of generalizations need to be operationally defined; the investigator is probably in a better position than anyone else to indicate their meaning.
11) The translation of factual data into value judgments should be made with care. Factual materials verified from more than a single point of view are ordinarily more readily translated into action patterns than the half-truths that may arise out of fragmentary research. In general we accept someone else's inferences only if we agree with his semantic and statistical postulates, his sources of information, and methods of induction.

DEVELOPMENT OF AN INVESTIGATIONAL DESIGN

Educational research is a complex activity; only through the most meticulous specifications can the many factors that need to be kept in mind be controlled at the proper time. In planning a building, such as one's home, years sometimes are spent collecting and organizing pertinent data into a plan. Educational research and appraisal are infinitely more complex than house planning. Improved research will result only when we plan with the utmost care.

In its final form the plan should be sufficiently complete and precise that all workers engaged in the project can follow the plan without doubt about what is to be done. A good plan is also so

drawn that other investigators removed in time and location can repeat the investigation.

Questions to be answered. Many important prior decisions should be made. Among these decisions will be answers to questions such as the following:

1) To what kinds of question are answers to be sought?
2) Is the goal sought a description and appraisal of some immediately at hand group of objects or persons or of the development of inferences about populations from carefully drawn samples?
3) Is the research to be of the carefully restricted and controlled single variable kind or of the more inclusive multi-variable, statistically controlled field type?

Although investigations differ in detail, there are many common questions to which they seek answers. The more frequent of these are the following:

1) What is the status of something?
2) Is this status adequate?
3) How has the situation under investigation come to be what it is?
4) How are the aspects of a situation, object, or phenomenon interrelated?
5) Are the problem-solving techniques employed in solving different kinds of problem adequate?
6) Can reliable predictions be made concerning some future status?
7) How can present practice be improved?

Questions such as these may be answered relative to almost all aspects of the educational program: the personnel, the socio-physical setting for pupil growth and achievement, the curriculum, materials of instruction, administrative practices and organization, teaching methods and guidance, financial support, and school-community relationships.

Descriptive versus inferential research. In teaching a class or administering a school system, one's interest may be merely in the present status of some aspect of the educational program close at hand and the complex interplay of forces that bring it about. One's immediate concern may not be with other schools or populations but with the assignment immediately at hand. The problem is one of accurately describing and appraising the materials, processes, conditions, and pupil changes that constitute one's

immediate responsibility. One's purpose is to evaluate a plan of action.

In many instances, however, one's concern is not with the group at hand but with future groups or more inclusive current groups. To make valid and reliable inferences about either, one must concern himself with the methods of sampling. The goals of science are description, explanation, and prediction. To predict one must obviously concern himself not merely with the immediate group but with future groups or more inclusive current groups according to one's purpose. While we may wish to explore and describe the characteristics and their interrelationships for some immediate group, we may also wish to describe by inference present and future populations. The latter purpose can be attained only when careful attention has been given to sampling techniques.

Inferential research is important in that it carries one beyond the immediate situation; descriptive research is important in that it may serve not only our present needs but that it may establish a basis for important inferences in sampling research.

Descriptive research may include the study of relationships. Descriptive research may include the study of relationships among aspects of the educational program such as materials, processes, conditions, and outcomes. Correlational studies, in particular, may be of the descriptive type. Even the classical single variable type of experiment may seek to ascertain the nature of the relationships on a descriptive level. Although none of these may involve sampling, some investigators will prefer to differentiate between descriptive status studies, descriptive correlational studies, and descriptive experimental studies and other nonsampling techniques.

Descriptive status research may be very important. Many persons regard this very important type of fact-finding as less worthy than that of drawing inferences from samples about populations. Both descriptive and inferential research are important. Whether human insight and ingenuity are more in evidence in discovering and validating important facts and relationships about immediate phenomena than the study of inclusive populations through the use of appropriate mathematical models and sampling techniques is probably not an issue of great importance. Both types of data and approaches to research and appraisal are necessary.

Studies of the current and contemporary versus historical foundations. Educational research and appraisal as currently conducted draw heavily upon present conditions and events. To understand many current phenomena it is frequently important, however, to understand how they have come to be. Studies of the past

extend the intellectual horizon by indicating how things have originated in remote times and places and how these may have affected the present. They also help in differentiating between the important and the superficial, and the continuing from the passing. Attitudes, values, and emotionalized forms of behavior almost always have a history; not to understand this history is not to understand the particular behavior under investigation.

Prediction versus explanation. Reliable predictions and control over educational outcomes arise only when the components of phenomena and their relationships are understood and the understandings derived from subgroups adequately generalized through accepted sampling procedures. Understandings of developmental sequences and historical foundations frequently provide much needed information for stable predictions and controls.

Artificially controlled laboratory research versus field research. There are other broadly different approaches to research. One involves the single variable laboratory approach to research and the other the multivariable field approach. Both approaches make valuable contributions to human understanding. Emphasis has been placed in this volume, however, upon the more complex types of field research and appraisal. For the conventional classical approaches to research and appraisal the reader is referred to a number of books on this subject. Good, Barr, and Scates'[1] *Methodology of Educational Research* and Westaway's[2] *Scientific Method* are helpful. Dewey[3] in his *Sources of a Science of Education* raises some of the issues involved in laboratory versus field research. The investigator's choice of research procedures will be governed by needs of the situation.

Other factors to be considered in the choice of investigational design. Objectivity is an important concern of the research worker. It is intimately associated with the development of trustworthy evidence. In noting evidence it is frequently worth while to ask the question: will these data divide the audience, so to speak, or will they lead to acceptance and agreement? Objectivity is a goal that should be always before us. A closely related concept is reliability. Every measurement is subject to error; the standard or probable error should therefore always be stated. The choice of calculational procedure in inferential research is no less impor-

[1] Carter V. Good, A. S. Barr, and Douglas Scates, *Methodology of Educational Research* (New York: D. Appleton-Century Co., 1936).

[2] F. W. Westaway, *Scientific Method: Its Philosophical Basis and Modes of Application* (London: Blackie and Son, 1937).

[3] John Dewey, *Sources of a Science of Education* (New York: Horace Liveright, 1929).

tant. The investigational design should be so planned as to permit a probability statement about the likelihood of each generalization.

Need for more inclusive long-time field studies. Emphasis has been placed in this volume upon the complexity of educational appraisal and research and the necessity for a multiple approach —multiple with respect to investigators, data-gathering devices, situations and persons, research design, and analytical techniques. To get a better picture of what is taking place, a broader view of the learning-teaching situation must be developed and introduced into investigational designs. More factors, over longer periods of time, should be studied with more persons with proper recognition of their settings and social foundations. Everyday activities of the school involve continuous, long-time, and inclusive searchings for the truth.

The need for an adequate record system. Continuous long-time inclusive research necessitates a good system of records that will provide reliable data on many aspects of the educational program during long periods of time. In addition to the data included in the typical record system there will need to be more information about each responding individual, about the conditions under which responses were obtained, and socio-physical context from which they are taken.

We wish to turn now to materials, illustrative of these more important characteristics of good research. Different approaches to research have been chosen to emphasize the more important components of properly planned investigations. The first illustration is a developmental study in which personal orientation predominates. The investigator's concern was with this personal orientation. The techniques used could be more generally employed in all research.

DEVELOPMENTAL STUDIES OF INDIVIDUAL SCHOOL CHILDREN

Through the use of the case study method emphasis is placed upon the personal orientation of data. We need, however, a long view of how individuals become what they are at any particular point in their developmental history, if we are to survey progress and study interrelationships. With such developmental data for all the individuals of carefully described groups within some well designed investigational pattern, much more meaningful relationships than those obtained from cross-sectional research may be

had. Such individualized developmental data, socially oriented, should provide the basis for more trustworthy generalizations than those ordinarily secured from current approaches. The literature of education is rich in developmental studies of the restricted type, but none seems to encompass the scope suggested here. Any one of a large number of such studies [1,2] might be chosen however to illustrate the kinds of data that might be collected in a more complete analysis. We have chosen *Development in Adolescence* by Harold E. Jones [3] to illustrate the type of data needed about persons as one part (aspect) of a more inclusive study of educational programs.

The Jones Developmental Study. Many aspects of development were studied by Jones and his associates:

I. *John at home:*
   A. Community and neighborhood
   B. John's mother
   C. John's father
II. *John at school:*
   A. Elementary school
   B. Junior high school
   C. Senior high school
III. *John as seen by his teachers and classmates:*
   A. Scholarship record
   B. Teacher's comments
   C. Classmates' opinions
   D. Group structures
IV. *John as a member of social groups:*
   A. Boys' groups
   B. Behavior among boys and girls
V. *Physical development:*
   A. Health record
   B. Physical growth
   C. Skeletal maturing
   D. Growth curves
   E. Physiological changes
VI. *Motor and mental abilities:*
   A. Strength records
   B. Physical abilities
   C. Manual abilities
   D. Mental abilities

[1] Arnold Gesell and Catherine C. Amatruda, *The Embryology of Behavior: The Beginnings of the Human Mind* (New York: Harper, 1945).

[2] Arnold Gesell, Frances L. Illg, Louise B. Ames, and Glenna E. Bullis, *The Child from Five to Ten* (New York: Harper, 1946).

[3] Harold E. Jones, *Development in Adolescence* (New York: Appleton-Century-Crofts, 1943).

- E. Achievement tests
- F. Learning ability

VII. *Interests and attitudes:*
- A. Group changes
- B. Areas of interest
- C. Religious beliefs
- D. Social attitudes

VIII. *Underlying tendencies:*
- A. Drive patterns
- B. Projective materials
- C. Voice records
- D. Rorschach records
- E. Emotional trends

IX. *As John saw himself:*
- A. A personal-social inventory:
  1) deficiencies
  2) aspirations
  3) family relationships
  4) attitudes toward school
  5) adjustment
- B. John as a judge of himself and others:
  1) association of traits
  2) omissions of items
  3) conformity with group opinion

X. *Struggle for maturity (a summary statement).*

Many data-gathering devices were employed in the study: school records, teachers' reports, classmates' opinions expressed in the form of sociograms and a reputation test, observations of behavior and social relationships leading to anecdotal records and social ratings, health examinations, body photographs, x-ray assessments, physical measurements, basal metabolism studies, strength records, physical performance and ability tests, manual dexterity tests, intelligence tests, achievement tests, learning ability tests, activity inventories, interest inventories, beliefs and opinion ballots, projective devices, voice records, photographic records of psycho-galvanic reactions, and a personal social inventory to obtain material relating to John's estimate of himself in many relations.

Jones' account of the study follows: The report is based on records obtained for an individual during a period of seven years. This person was selected from a grade group of eighty boys, in a growth study consisting originally of approximately 200 boys and girls.[1] It is perhaps difficult to say why "John Sanders" was chosen

[1] The general plan of the study is described in the *Journal of Educational Research,* 1936, 31:561–567.

for this presentation, rather than any other of his classmates. Others might have served the general purpose equally well, although some special interest attaches to this case because he presents a number of problems which are of common occurrence in contemporary urban culture. In the case of John, these problems of adjustment are less remarkable for severity than for variety. John has been handicapped by unhappy relationships within his family; economic stress; ill health; visual defects; an inferior physique; delayed maturity; a certain obtuseness in social contacts; lack of athletic abilities; and lack of ability to win goals which he has most desired in connection with a strong drive for popularity and social esteem. Under this heavy accumulation of handicaps, his school career has been notable for a cycle of personal difficulties followed by some degree of success and effective adjustment. Reviewing the record of John Sanders from the sixth grade to college, one cannot help being impressed by the amount of idiosyncrasy to be observed within a "normal" range, and by the complexity of problems which can be faced and even to some extent surmounted within a social structure that has done little to provide systematic support or understanding.

The author concludes as follows:

> John Sanders was a boy with an extraordinary accumulation of personal handicaps: physical, social, emotional, economic. He was unsupported by any special sense of security in his family; unaided by any special gift of intelligence or by any special insights on his part or on the part of his teachers. He reached a low point in adjustment, but he did not remain there. The greater personal stability and the more adequate social relationships he achieved in the last year of high school were carried forward during college. His college years also brought a successful record in courses and in an enterprising variety of outside activities. So marked an upturn in John's personal fortunes is evidence not only of the toughness of the human organism but also of the slow, complex ways in which nature and culture may come into adaptation.[1]

The report cited above is for a single case. In many instances one's concern will be with single cases but when the purpose is to derive trustworthy generalizations about a well-defined population, a carefully prepared plan of sampling should be followed. One of the especially difficult problems of all such research pertains to the selection of the aspects of the phenomenon to be studied. The early writers in this field recommended that investigators collect all possible information with the fewest possible precon-

[1] Harold E. Jones and others, *Development in Adolescence* (New York: Appleton-Century-Crofts, 1943).

ceived ideas and let the facts lead wherever they might. In hypothetical thinking one does not, however, collect all possible data but only those believed to be pertinent to the investigation, or those shown to be so by previous investigations. No investigation, even of individuals, is better than the information-gathering devices employed in collecting the data for study and analysis.

COMPLEX STUDIES OF INTERRELATIONSHIP

There are in the literature a number of cross-sectional studies of the complex interrelationship that surrounds the learning-teaching situation. To emphasize certain aspects of such studies reference is made to studies of Rostker,[1] La Duke,[2] and Lins.[3] All of these studies are of the more complex sort and could have been developmental in character if appropriate data were available.

La Duke and Rostker were particularly anxious in the planning of their research that the teachers and the investigators agree on pupil objectives. After outlining the purpose of his study and describing the investigational design, Rostker reports the directions to the teachers, and the pupil objectives, as follows:

> The plan proposed and executed was: (1) to measure pupil performance near the beginning of the school year and near the close of the school year so as to obtain long-time pupil changes occurring over approximately six months; (2) to measure pupil performance just prior to the teaching of and immediately after the teaching of two three-week units of work in the general field of citizenship—one of these units to be given in the fall of the school year, the other unit in the spring of the same school year—so that pupil changes on two short-units of work would be obtained; and (3) various measures and rating scales were applied to the teachers, preferably in the fall of the school year, with the exception of those tests taken by both teachers and pupils which would be given concurrently.
>
> Several weeks before any teaching of the units occurred, each teacher was sent a letter in which were stated the dates when testing was to begin, a statement of the topics, and the general objectives in terms of desired goals for which the teachers, teaching the first unit, "Safeguarding Public Health," were to strive.
>
> The topics to be included in the first unit were:

[1] L. E. Rostker, "The measurement of teaching ability," *Journal of Experimental Education,* 1945, 14:6–51.

[2] Charles V. La Duke, "The measurement of teaching ability," *Journal of Experimental Education,* 1945, 14:75–100.

[3] Leo Joseph Lins, "The prediction of teaching efficiency," *Journal of Experimental Education,* 1946, 15:2–60.

1) Securing pure air, food and sunshine.
2) Disposing of wastes.
3) Providing desirable housing.
4) Caring for the physically and mentally sick.
5) Recreational opportunities.

The goals toward which the teachers were to direct their instruction for this unit were: (1) "to acquire the kinds and amounts of information essential to the understanding of the problems and issues involved in safeguarding public health"; (2) "to develop skill in forming judgments about this subject"; (3) "to develop desirable attitudes relative to safeguarding public health"; and (4) "to lead the pupils, individually and co-operatively, to some positive action relative to safeguarding public health."

Following the close of the first unit, several months intervened in which the teachers resumed their normal course of study. In the Spring of the same school year (early March, 1937), the participating teachers were informed that on certain dates they would be requested to begin teaching a three-week unit on "Community Planning."

The topics covered in this unit were: (1) the layout of streets; (2) the building zones; (3) beautifying the community; (4) keeping the community clean; and (5) recreational facilities.

The goals of instruction sought for this unit were the same as those sought in the first unit with the exception of differences in subject matter.

La Duke describes the development of a course outline to be used by all teachers participating in his investigation as follows:

The outline of the course of study for Community Living, composed of eight units, was prepared by members of the State Department of Public Instruction, a committee of teachers of the social studies who volunteered to assist in the planning of the series, and members of the radio project research staff. The attempt was made to portray the progressive development of the political and social organization of our democratic society. The child's responsibility in the functioning of democratic society was to be emphasized.

The following outline and schedule was developed:

Unit I. Your Family, Home, and Community.
September 26. You and your family.
October 3. You and your home.
October 10. Your home and your community.
Unit II. How the Community Serves You.
October 17. Safe highways.
October 24. Protection of life and property.
October 31. Education—your opportunity.

November 7. Recreation—a new community service.
November 14. Health—a community problem.

Unit III. How the State Serves You.
November 21. Conservation of natural resources.
November 28. State protection for producer and consumer.
December 5. Services of the state university.
December 12. The Wisconsin Social Security Law.

Unit IV. How the National Government Serves You.
January 2. Uncle Sam carries the mail.
January 9. Government research serves your home.
January 16. Government regulation—labor, communication, etc.
January 23. Uncle Sam cares for the unemployed.

Unit V. How Your Government is Organized and Supported.
February 6. Managing your local government.
February 13. Managing your state government.
February 20. Managing your national government.
February 27. Paying the bill together—taxation.

Unit VI. Making a Living in Your Community.
March 6. Making a living in the country—agriculture.
March 13. Making a living in the city—Wisconsin industries.
March 20. Workers' problems in town and country.
March 27. Buying and selling together—cooperatives.

Unit VII. Social and Political Groups.
April 3. Nationality groups in Wisconsin.
April 10. Social life in your community.
April 17. Political parties.
April 24. Your part in democracy.

Unit VIII. Your Community and World Society.
May 1. Your state and world markets.
May 8. Your country and world problems.
May 15. Your part in the world community.

The objectives of the course were formulated by members of the research staff who attended the Progressive Education Workshop held in Bronxville, New York, during June and July, 1938. Leaders in the field of social studies who were in attendance at the conference assisted in a large measure in the formation of these objectives.

The following specific objectives for Unit III, How the State Serves You and Your Community, are illustrative of the objectives developed:

SPECIFIC OBJECTIVES

A. Functional Information:

1) To develop an understanding of how the state serves its citizens and their many interests and needs.
2) To indicate the ways in which the services of the state are related to the services of local government.

3) To indicate the community needs which make state services necessary.
4) To indicate the manner in which the state provides for the conservation of its natural resources.
5) To indicate the ways in which the state attempts to protect the interests of the producer and consumer.
6) To indicate some of the services provided by the state university for the citizens of the state.
7) To indicate the manner in which the state provides for social welfare—social security.
8) To indicate the deficiencies in the service of the state.

B. *Interests:*
1) To develop an interest in the functions of the state government.
2) To develop an interest in the ways the state protects our natural resources.
3) To become interested in the ways in which the state protects the producer and the consumer.
4) To become interested in the services provided by the state university.
5) To become interested in the problem of providing for the social welfare of its citizens.

C. *Appreciations:*
1) To recognize the value of the services of the state to the individual and to the community.
2) To recognize the importance of conserving our natural resources.
3) To develop a recognition of the state's responsibility for social welfare.
4) To recognize the importance of the services provided by the state university.
5) To develop an appreciation of the interdependence of the producer and the consumer.
6) To develop an appreciation of the individual contributions to the growth of state services.

D. *Attitudes:*
1) To develop an attitude of cooperation toward state services.
2) To develop an attitude of concern with regard to state activities.
3) To develop a sense of responsibility in improving the services of the state.
4) To promote an attitude of concern regarding the conservation of natural resources.
5) To develop an attitude of consideration of the rights and needs of the different interest groups within the state in regard to governmental services.

Each of these investigators obtained some teacher participation in the planning of these investigations but not the full and extended co-operation currently recommended wherein teachers, pupils, and members of the community participate in educational planning. Their plans provided for complete information about what was to be done, even though there was only limited participation in the planning. Large amounts of data were collected about pupils and teachers. Rostker's list of tests and other data gathering devices applied to the pupils and teachers is given below:

A. Measuring devices applied to pupils:

1) Kuhlmann-Anderson Intelligence Test, Fourth Edition, Grades VII–VIII, 1933.
2) Traxler Silent Reading Test, Form I, Grades 7 to 10.
3) Sims Score Card for Socio-Economic Status, Form C.

B. Measuring devices applied to the teachers, tests:

1) The American Council on Education Psychological Examination for College Freshmen, 1936 edition.
2) The Teachers College Psychological Examination, 1934 edition.
3) The American Council Civics and Government Test, Form B.
4) Social Attitudes of Secondary School Teachers.
5) A Scale for Measuring Attitude Toward Teachers and the Teaching Profession (by Tressa C. Yeager).
6) The Morris Trait Index L (by Elizabeth H. Morris).
7) Orientation Test Concerning Fundamental Aims of Education, 1935 Revision (by Alfred S. Leverenz and Harry C. Steinmetz).
8) Personality Inventory (by Robert G. Bernreuter).
9) Social Adjustment Inventory (Sapich Edition) (by J. N. Washburne).
10) Stanford Educational Aptitudes Test (by Milton B. Jensen).
11) Test of Teaching Problems (by T. L. Torgerson).
12) Theory and Practice of Mental Hygiene (by T. L. Torgerson).
13) Abilities to Organize Research Material (by J. W. Wrightstone).
14) Test on Community Planning, Form A.
15) Safeguarding Public Health, Form A.

C. Measuring devices applied to the teaching, rating scales:

1. Almy-Sorenson Rating Scale for Teachers (by H. C. Almy and Herbert Sorenson).
2. Michigan Educational Association Teacher Rating Scale (by the Michigan Education Association).
3. Diagnostic Teacher Rating Scale of Instructional Activities (by T. L. Torgerson).

As a result of careful analysis of the data collected about the pupils a consistent picture of their accomplishment and growth during the experimental period was attempted. The means and standard deviations, for each measuring device and for each class were calculated, as well as their intercorrelations. These data furnished a wealth of material showing unevenness in the achievement of different pupils and groups of pupils for different tests. Similar analytical procedures were applied to the data collected about each teacher.

The intercorrelations between pupil growth and achievement scores, predicted scores, measured teacher abilities, and teacher efficiency were also calculated. Some of the correlations are given in Table 42. Many consistencies and discrepancies were discovered in these data. Whether these consistencies and inconsistencies are important will need to be ascertained from such further differential analysis of learning and teaching as one might obtain from continuing research in this field. For the conditions under which this study was made, Rostker obtained a multiple correlation of .85 between fourteen measures of teacher ability and a composite criterion of pupil growth and achievement. Under a different set of conditions La Duke obtained a multiple correlation of .80 between four selected teacher measures and his criterion.

Much effort was expended in these studies upon the development of criteria of teaching efficiency. In addition to measures of pupil growth and achievement used by Rostker and La Duke to measure teacher effectiveness, Lins employed an elaborate system of supervisor and pupil ratings. For the supervisory rating a composite of five ratings was secured through the use of the Wisconsin adaptation of the M-Blank of the evaluative criteria. Three of the ratings were made by members of the Department of Education, who visited each teacher; one by a member of the State Department of Public Instruction; and one by the principal or superintendent of schools under whom the teacher did her teaching. Preparatory to making these ratings the raters spent the greater part of one full year in discussing the criteria to be used and the method of collecting data. The reliability of the supervisory ratings criterion as determined by the chance halves method and the Spearman-Brown prophecy formula was .86.

The criterion of pupil evaluation was based upon the ratings of a number of pupils (usually five or six) chosen from those taught by each teacher. The directions for these ratings are given below:

> You are now in several different classes or courses taught by several different teachers. I am visiting Miss X and she is one of your teachers.

You can help us train teachers at the University of Wisconsin by telling us how good she is or bad she is, whichever is the case. What you write on your paper is strictly confidential; it will be seen by no one else but me. Now think over the list of different teachers from whom you are now receiving instruction. Write down the number of different persons from whom you are now receiving instruction. If Miss X is the best of these several instructors write down a number 1 and draw a circle around it; if she is second best, that is if you can think of one better, write down a figure 2 and draw a circle around it; and so on. If she is the poorest simply write the word last or poorest. If I may have your paper now I will put it here in this large envelope and seal it. I appreciate your assistance.

Some pupils had three teachers, some four, and some five. The ranks were all transmitted to a five-step scale with values of 5, best; 4, next best; et cetera. The several ratings for each teacher were then averaged for a single composite score. This last named composite score is that used as the criterion of teaching efficiency employing pupil evaluations.

Among the interesting findings of the Lins study were the low coefficients of correlations obtained between these several composite measures of teaching efficiency. These were as follows:

1) The correlation between the pupil gain criterion and the supervisory ratings was .19.
2) The correlation between the pupil gain criterion and pupil evaluation of teaching was .05.
3) The correlation between pupil evaluation of teaching and supervisory ratings was .28.

Obviously notwithstanding the care with which the several composite measures of teaching efficiency were developed, the intercorrelations among these measures were not high, indicating need for further research.

The three studies described above have been chosen to illustrate complex cross-sectional investigations of educational operations that might be extended over long periods of time. As detailed as these studies were, only a few of the many aspects of education that one might study were investigated. If one distinguishes between descriptive and inferential research, these studies are of the nonsampling sort. Special care was exercised by these investigators to indicate the purposes and conditions of these investigations and the precise character of the data-gathering devices employed. The statistical analysis was reasonably adequate for their purposes. Nevertheless these investigations might have been considerably improved: by a more explicit delineation of the many assumptions made throughout the investigation; by a better description of the

Table 42. Intercorrelations Among the Teacher

		T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
T 1	Wrightstone Research	1.00	.49	.59	.39	.47	.50	.32	.33	.65	.44	.01	.50	.07	.34
T 2	American Council Psychological Exam.		1.00	.57	.31	.38	.80	.15	.34	.60	.04	−.22	.35	−.20	.50
T 3	Hartman Social Attitudes			1.00	.33	.41	.42	.14	.28	.55	.46	−.30	.50	.21	.34
T 4	Yeager Attitude Toward Teaching				1.00	−.08	.12	.09	.01	.32	.02	.07	.14	.02	−.03
T 5	Torgerson Mental Hygiene					1.00	.34	.31	.23	.52	.30	−.13	.52	−.21	.31
T 6	Teachers College Psychological Exam.						1.00	−.02	.42	.58	.44	−.03	.30	.04	.27
T 7	Community Planning Test							1.00	.35	.09	.04	−.10	.19	.06	.09
T 8	Health (unit) Test								1.00	.21	−.05	−.18	−.08	.24	.04
T 9	American Government & Civics									1.00	.50	.03	.54	−.06	.32
T10	Torgerson Teacher Rating Scale—I										1.00	−.27	.55	−.33	.87
T11	Bernreuter Personality Inventory—Bn											1.00	−.02	.67	−.38
T12	Orientation Test												1.00	−.17	.41
T13	Bernreuter Personality Inventory—Fc													1.00	−.31
T14	Almy-Sorenson—I														1.00
T15	Bernreuter Personality Inventory—Bd														
T16	Michigan Teacher Rating Scale—I														
T17	Morris Trait Index L														
T18	Bernreuter Personality Inventory—Bs														
T19	Michigan Teacher Rating Scale—S														
T20	Bernreuter Personality Inventory—Fs														
T21	Washburne Social Adjustment Inventory														
T22	Teachers Problems														
T23	Stanford Educational Aptitude Test—T—A														
T24	Stanford Educational Aptitude Test—A—R														
T25	Almy-Sorenson Teacher Rating Scale—S														
T26	Stanford Educational Aptitude Test—T—R														
T27	Torgerson Teacher Rating Scale—S														
C 1	Composite of Unit Tests (raw score gain)														
C 2	Composite of Wrightstone Tests (raw score gain)														
C 3	Composite of Hill Tests (raw score gain)														
C 4	Composite of Unit, Hill, & Wrightstone (raw score gain)														
C 5	Composite of Unit Tests (adjusted gain)														
C 6	Composite of Wrightstone Tests (adjusted gain)														
C 7	Composite of Hill Tests (adjusted gain)														
C 8	Composite of Unit, Hill, & Wrightstone (adjusted gain)														

Measures and the Eight Criteria (N = 24)

T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	T25	T26	T27	C1	C2	C3	C4	C5	C6	C7	C8
.05	.25	.40	.14	.18	.21	-.02	-.08	-.22	.08	.19	.02	.13	.31	.44	.43	.49	.34	.49	.55	.58
.16	.57	.47	.04	.47	.03	.25	.11	.02	-.06	.31	-.10	.38	.26	.56	.40	.58	.25	.58	.29	.57
.25	.41	.31	.36	.31	.05	.18	.05	.03	-.23	.21	-.32	.10	.26	.49	.21	.50	.29	.49	.29	.52
.15	.03	-.06	.15	-.14	.19	-.14	-.53	-.22	.08	-.34	-.02	.24	-.03	.47	.43	.45	-.03	.40	.47	.45
-.19	.27	.53	.18	.13	.16	.12	.26	.04	.36	.18	.36	.28	.50	.40	.27	.46	.54	.38	.34	.45
-.22	.33	.24	-.27	.44	-.16	.32	.19	.02	-.09	.27	.20	.32	.32	.30	.47	.37	.31	.33	.42	.40
.18	-.08	.10	.38	.08	.20	-.18	.07	.04	.41	.24	.36	.02	.29	.42	.14	.44	.29	.37	.06	.39
-.03	.01	-.06	-.06	.07	-.31	.13	.03	.14	-.24	-.02	-.25	-.04	.17	.31	.11	.31	.19	.36	.17	.37
-.13	.35	.48	.16	.19	.24	.02	-.17	-.41	-.09	.04	-.12	.11	.29	.30	.34	.37	.31	.33	.40	.36
.30	.86	.46	.27	.78	.07	.22	.39	-.12	-.04	.62	-.03	.66	.33	.16	.18	.19	.32	.30	.20	.34
-.81	-.39	.08	-.50	-.32	.50	-.58	-.28	-.41	.14	-.21	.13	-.09	-.20	-.27	-.11	-.28	-.20	-.32	-.08	-.31
-.06	.32	.23	.52	.22	.21	-.13	.05	-.13	.19	.27	.09	.29	.35	.24	.23	.27	.37	.32	.24	.30
-.49	-.38	-.22	-.50	-.07	.24	-.47	-.24	-.32	-.13	-.09	-.12	-.04	-.20	-.29	-.07	-.29	-.21	-.27	-.09	-.27
.35	.91	.42	.36	.81	.03	.16	.42	-.14	.04	.62	.14	.71	.34	.11	.10	.15	.34	.22	.19	.26
1.00	.38	.05	.65	.28	-.16	.26	.08	.19	-.07	.12	.00	.04	-.09	.34	-.16	.27	-.09	.34	-.12	.25
	1.00	.42	.32	.81	.06	.25	.44	.02	-.15	.62	-.09	.70	.31	.12	.12	.16	.31	.18	.21	.23
		1.00	.15	.15	.47	-.04	.09	-.37	.08	.13	.43	.13	.01	.30	-.03	.26	.03	.24	.00	.20
			1.00	.27	.23	.05	-.02	.10	.13	.22	.11	.12	-.03	.24	-.20	.23	.00	.27	-.11	.20
				1.00	.02	.08	.47	-.02	.01	.84	.00	.81	.29	.02	.10	.07	.27	.07	.08	.15
					1.00	-.56	-.21	-.49	.22	.08	.29	.04	-.32	.09	-.27	-.09	-.30	-.05	-.21	-.13
						1.00	.43	.52	-.32	-.06	-.30	.13	.17	.12	.07	.14	.16	.12	-.07	.13
							1.00	.35	-.08	.50	.03	.52	.44	-.03	-.05	.04	.42	.06	-.11	.11
								1.00	-.33	.16	-.30	.17	.14	.12	-.01	.13	.14	.06	-.06	.10
									1.00	.05	.90	-.01	.10	.05	.18	.11	.10	.00	.16	.04
										1.00	.05	.81	.30	-.15	.03	-.08	.28	-.12	.05	-.02
											1.00	.03	.03	.07	.09	.08	.03	.01	.07	.02
												1.00	.31	-.14	.05	-.06	.33	-.09	.02	-.01
													1.00	.25	.69	.44	.99	.36	.60	.55
														1.00	.46	.98	.26	.95	.26	.91
															1.00	.63	.68	.49	.84	.66
																1.00	.45	.94	.43	.94
																	1.00	.38	.61	.56
																		1.00	.35	.96
																			1.00	.57
																				1.00

socio-educational setting and background of the practices and achievements observed; and by a more critical technique for establishing cause and effect relationships. These investigations are correlational and not experimental investigations. Inasmuch as these studies were descriptive, the generalizations were very appropriately limited to the persons studied, under the conditions observed, and the particular data-gathering devices employed.

A community-wide project involving a study of social foundations. A third study has been chosen for two purposes: (1) to illustrate community wide planning; and (2) to illustrate how sociological data may be correlated with educational data. The study was part of the Wisconsin small high school study.[1]

The responsibility for the planning of this study, the collection and analysis of the data, and the action program to follow was placed in the hands of a large volunteer committee composed of approximately fifty members. The committee selected seven communities to co-operate actively with it in its work. These communities were: Blair, Cambridge, Campbellsport, Hancock, Johnson Creek, Winneconne, and Wonewoc. The communities represented a wide geographic coverage of the state without too much travel. Seven criteria were employed in choosing the schools:

1) The school must be within the size limits set.
2) The school must be within the middle 50 per cent of its size group in pupil-teacher ratio.
3) The school must be within the middle 50 per cent of its size group in annual instructional cost per pupil.
4) The community must be definitely rural in character.
5) Fifty per cent of the high school pupils must be living on the farm while attending school.
6) Any school whose buildings were hopelessly inadequate or distinctly superior was not eligible.
7) The final sample must include all the major types of agriculture found in Wisconsin.

To secure local co-operation, a first conference was held with the high school principal and the county superintendent jointly to explain the project. The conference was followed by a meeting with the local school board and representative citizens interested in better schools. With the approval of the board the program was explained to the teachers. Co-operative agreements were signed with the several communities to cover a five-year period.

The specific objectives of the study were stated as follows:

[1] C. E. Ragsdale, and others, "Adventures in rural education: a three-year report," *Journal of Experimental Education*, 1944, 12:145–348.

1) To improve rural education in the state at large as a result of experiences with a limited number of communities.
2) To assist the co-operating communities to plan and operate their schools with special reference to their needs as rural areas.
3) To assist especially in improving rural secondary education.
4) To encourage the planning of a unified twelve years of rural education.
5) To try out plans for better relating rural schools to the resources and activities of the community.
6) To study the problems of and foster improvement in teacher education for rural elementary and secondary schools.
7) To improve the attractiveness of teaching in rural elementary and secondary schools and to provide for continued improvement in service of teachers.

To carry out the work, a series of subcommittees were established: a central executive committee; committees in the co-operating communities; and special committees on report cards, teacher training, English, health (recreation, and physical education), curriculum, and correspondence study.

The local schools and committee members were visited to explain functions and to develop procedures for holding conferences and carrying forward the program of activities. The first request of the central committee was that the principal, the teaching staff, and the school board think over their local situation and formulate two or three problems for study. These problems ordinarily centered around the curriculum, the need for a unified educational program, the teaching personnel, and means of taking advantage of the strength of rural life.

The data-gathering devices were principally those of conference and record study. The conclusions were the consensus of participating members.

The report on Cambridge covers the following topics:

I. The Community and Its Schools:
- A. The village and its farm community.
- B. School finances and equipment.
- C. The school population.
- D. The teaching staff.
- E. The instructional program.

II. New Ventures in Community Education:
- A. Introducing home economics.
- B. A comprehensive community survey.
- C. Discovering of school problems.
- D. A school community dinner (presentation by president of

senior class of a summary of improvements, some 46 items, in the Cambridge High School).

E. Cambridge High School noon lunch program.
F. Cambridge noon activity council.
G. Changes in the Cambridge High School library (some 52 items are listed).

The report for the Wonewoc community project covers a similar list of topics:

I. The Community and Its Schools:
  A. Geographic features of the community.
  B. Settlement and early growth.
  C. Industry, commerce, agriculture, and people.
  D. The school population.
  E. The teaching staff.
  F. The instructional program.

II. New Ventures in Community Education:
  A. Background for co-operative work.
  B. Curriculum changes.
  C. Correlation of English and social studies (a list of units is given).
  D. Pupil forum.
  E. Community survey by seniors.
  F. Survey of job opportunities.
  G. Occupations of graduates.
  H. Health education and service.
  I. Hot lunch programs.
  J. Improved library service.
  K. Use of motion pictures.
  L. Improved school records.
  M. New school bus routes.
  N. A new kindergarten.
  O. Improved elementary education (a list of science units is given).
  P. Wonewoc Tri-county Forum (a list of topics is given).
  Q. The community builds a playground.

Plans, actions, and results are summarized under these several headings. The strength of this approach will be found in the planning; in widespread participation of pupils, teachers, and community; and in the sociological setting provided for the data collected. There is less quantitative data than one would desire, but this is not a characteristic of this approach, only a characteristic of this particular study.

It is difficult to do justice to this and other investigations from the brief excerpts reproduced. The last-named investigation is of

the nonsampling descriptive type. In general the data were gained from observation and interview and recorded in verbal symbols. When generalizations were reached, they were in the main those arising from a consensus of the participants. What the investigation lacked in objectivity it made up for in meaningfulness to the participants and the communities served. Other investigators working under different conditions might, however, wish to conduct long-range developmental studies with many objective data and efficiently designed experiments. These purposes were not, however, within the plans of this particular group of investigators.

SOCIAL DYNAMICS STUDIES

Behavior must always be considered from various points of view. Among these points of view is the social dynamics of the situation. Social dynamics as a field of study is concerned with the interaction between the various elements in the setting for action—particularly the interaction between the individual and his environment, especially his social environment. In social dynamics, various forces interact upon each other in a field. The individual's behavior in this field is determined by his internal state and the nature of the field forces. Much of the more conventional research would appear to ignore the situations out of which facts arise and treat those taken from different settings as equivalent or equal. Many valuable truths have been established through the manipulation of facts out of context. Behavior, however, arises from a combination of purposes, persons, and conditions; and its meaning varies with these conditions even though they may be given identical labels and tabulated as like objects, facts, or behavior. The developments in this area are too new and untested to predict their utility for research workers.

Lewin's formulation of the social dynamics theory. Lewin,[1] possibly more than any one other person, is responsible for our current formulation of topological and vector psychology. The assumptions of a field theory are (1) that behavior has to be derived from a totality of coexisting facts, and (2) that these coexisting facts have the character of a "dynamic field" in that the state of any part of this field depends on every other part of the field. According to the field theory behavior depends neither on the past nor on the future, but on the present field. Typical interaction studies re-

[1] Kurt Lewin, *The Conceptual Representation and the Measurement of Psychological Forces* (Duke University Press, Durham, N. C., 1938); and Kurt Lewin, *Formulation and Progress in Psychology* (University of Iowa Studies in Child Welfare, 1940, 16).

late to such matters as dominative and integrative behavior, social conflicts, frustration and regression, aggression and escape, competition and co-operation, and social cleavages.

Lippitt and White's[1] studies of the effects of different social climates. These authors report two studies, the first undertaken to develop investigational techniques and the second: (*a*) to study the effect upon individual and group behavior of three types of social atmosphere, labeled "democratic," "authoritarian," and "laissez faire"; (*b*) to study the transition from one atmosphere to another; and (*c*) to study the relationship between the child's home social climate and his adjustment to a particular club atmosphere provided by the experimental program. Authoritarian, democratic, and laissez-faire settings were operationally defined. In the authoritarian setting, for example, the leader determined all policy, assigned tasks and work companions, and appraised performance. Democratic and laissez faire were also defined in terms of performance.

The experimental plan provided for four clubs of five ten-year-old boys each roughly equated on patterns of interpersonal relationships, intellectual, physical, and socio-economic status, and personality characteristics.

Eight types of club records were kept, the most important of which were:

1) A quantitative running account of social interactions.
2) A minute by minute group-structure analysis.
3) An interpretive running account of strikingly significant member actions and changes in over-all group atmosphere.
4) Continuous stenographic records of all conversation.
5) Interview with each child.
6) Interview with the parents.
7) Talks with the teachers.

The reliability and validity of the data-gathering devices were given careful consideration. The data were analyzed with reference to the following categories:

1) Leader behavior.
2) Group morale.
3) Group and individual goals and achievement.

The data are statistical and descriptive in character. Some of the reasons for the comparatively low morale in authoritarian groups were given as:

[1] Ronald Lippitt and Ralph K. White, "The social climates of children's groups," in *Child Behavior and Development* edited by Roger G. Baker, Jocob S. Kounin, and Herbert F. Wright, McGraw-Hill Book Company, New York, 1943, pp. 485–508.

1) Restriction upon movement.
2) Frustration of the need for sociability.
3) Opposition to the leader and his goals.

Some of the reasons for group disruption in the laissez-faire groups were given as:

1) Restriction upon free movement.
2) Frustration arising from need of clearness of structure.
3) The vicious circle of frustration-aggression-frustration.

A careful study was made of goal setting in the three different social climates. It was concluded that co-operative group activity in the direction of group goals was impeded in authoritarian situation by a lack of identification with the group aims (as induced by the leader), and by social structure of the situation resulting from ego-centered in-group conflicts that hindered co-operative group progress.

The conclusions are supported by statistical and descriptive data. The studies of social dynamics emphasize among other things the complex interplay of forces from which human behavior arises. The context and the setting for behavior are also appropriately emphasized. Most of the social dynamic studies reported to date use observational and interview types of data-gathering devices. In recording what is observed one must sooner or later concern himself with qualities, aspects, and behavior, and their amounts. Most of the researches involving social dynamics have been of the descriptive type. As they are extended beyond exploratory studies of the immediate phenomenon they will doubtless concern themselves with sampling and efficient investigational designs for valid inferences about more inclusive groups and populations.

TREND STUDIES

Another type of complex developmental study which deserves wide use is the study of trends. Trends can be both desirable and undesirable. Those interested in educational planning and evaluation should be alert to both possibilities. Edwards,[1] for example, discusses in a series of articles the educational implications of population changes. His studies are concerned with such topics as the educational implications of declining fertility, the changing age structure of the population, differential fertility in relation to economic planes of living, the imbalance in educational load between

[1] Newton Edwards, "Educational implications of population changes," *Educational Forum,* 1946, 10:281–288, and "Educational implications of population changes," *Review of Educational Research,* 1946, 16:50–55.

regions and rural-urban communities, the reshuffling of the population, and federal aid.

The following excerpt taken from his statement relative to future prospects for population growth illustrate the factual nature of his data:

> A county by county analysis of reproduction rates for all the counties of the United States made by the writer and Herman G. Richey reveals that in some areas fertility is fully twice as great as in others. The great area of low fertility extends from southern New England southward to Maryland and spreads westward from New York and Pennsylvania, ending in southeastern Nebraska and western Kansas. The Pacific coast states constitute another major area of low birth rates. Fertility is strikingly higher in the Southern Appalachian-Ozark area, in the Cotton Belt of the Southeast, in parts of the Southwest, in the Rocky Mountain States, in parts of the Great Plains, in the cut over lands of the Great Lakes States, and in northern New England. When specific regions are compared, the number of children under five per thousand women 20 to 44 ranges from 341 in the Far West to 517 in the Southeast. Individual states exhibit even greater differences. For one group of states the number of children under five per thousand women 20 to 44 is: New York, 289; New Jersey, 294; and Connecticut, 312. For another group of states the corresponding ratios are New Mexico, 666; Utah, 593; South Carolina, 586; and Kentucky, 562.

Finally, he says:

> Careful analysis reveals that communities with the highest birth rates and the heaviest educational load are commonly the ones having the lowest planes of living and the weakest economic structure. In communities where the birth rate is low, the educational load light, the plane of living high, and the economic resources great, it is our policy to support education adequately for boys and girls who in all probability will not have enough children to replace themselves. In communities where the birth rate is high, the educational load heavy, the plane of living low, and the economic resources the most restricted we support education inadequately although at great effort. These conditions call for a modification of national educational policy. The ideal of equal access to education will remain an unrealized one for a long time unless the tax base for education is extended to include the entire nation. Moreover, the evidence points to the need of federal aid directly to worthy individuals as well as to the several states.

These are factual studies. Although we may differ among ourselves as to how such aid is to be administered, the data have been carefully collected and objectively treated. Further materials can be found on this subject in the many papers cited at the end of Ed-

wards' chapter, "Educational Implications of Population Changes" in the *Review of Educational Research.*[1]

Many aspects of the educational program lend themselves to trend studies: school costs, consolidation of schools, size of classes, teaching load, education of teachers, teachers' salaries, the content of textbooks, truancy, the composition of school boards, actions taken by school board, legislative controls over schools, court decisions, school buildings, and other tangible aspects of the educational program. Few of us realize the extent of factual data already available relative to these very important aspects of education. As a result of improved recording systems even more data might be made available for the study of these very important educational problems. Adequate sampling must always be the concern of those making trend studies. When a trend is noted it may be necessary to ascertain the character of the data, the extent to which the generalizations made may be based upon wishful thinking, and the character of the sampling.

SUMMARY

The purpose of this chapter has been to suggest how the many principles and techniques of educational appraisal and research may be combined into more adequate studies of the educational program. A multiple approach should enable the investigator to draw conclusions that will more adequately reflect a true picture of conditions than will the fragmentary studies limited to the minute aspects of the larger problem. No one investigator or group of investigators has as yet attempted the continuous, inclusive, complex sort of total research here envisaged. The developmental studies of the growth process constitute landmarks. The cross-sectional interrelationship studies as well as the experimental and statistical studies have made distinct contributions. Facts must also be considered in context both developmentally (as shown by social foundations research) and dynamically (as observed from the interplay of forces in the situations in which they originate). The data from these different approaches should be co-ordinated in such a way as to contribute substantially to the solution of important educational problems. While most research workers will not have the time, energy, or desire to conduct the comprehensive types of research suggested, recognition of the need for these comprehensive studies should nevertheless result in a basis for the intercommunication of ideas among research workers.

[1] Newton Edwards, "Educational implications of population changes," *Review of Educational Research,* 1946, 16:50–55.

APPENDIX A

Writing a Thesis[1]

Writing a thesis in education is an important activity and includes not only the preparation of a comprehensive report of one's investigation of a worth-while problem but all activities beginning with the initial selection of a problem, developing a procedure, collecting data, and summarizing results. One's ability to make a constructive contribution begins to increase at the moment when he first explores his interest in search of a problem which is yet in need of study and solution. When that state of developing a thesis arrives at which definite results are at hand, much of the work in the broad sense has been accomplished. Experience has shown that many students are unfamiliar with the significant steps that must be taken in the development of a thesis. The purpose of this discussion is to provide orientation for such students by considering some of the problems that arise in thesis writing.

Selecting a level and field of education. The student in search of a thesis problem will usually find it advantageous to decide whether he wishes to work in the level of elementary education, secondary education, or higher education. He may then concentrate attention upon some specific field within the chosen level such as teacher training, vocational guidance, physical education, educational measurement, character education, or methods of teaching.

Examination of original research. After some initial orientation, the student may direct his attention to original sources of information concerned with the chosen level and field of special interest. Original sources include [2] professional journals, monographs, and dissertations.

[1] Part of this material was published in the March, 1950, issue of the *Peabody Journal of Education.*

[2] If he is interested in elementary school problems, he may examine the *Elementary School Journal;* if in the secondary-school level, the *School Review;* if in the level of higher education, the *Journal of Higher Education.* Material of especial value for the research worker in education may be found in the *Review of Educational Research,* the *Journal of Educational Research,* the *Journal of Educational Psychology,* and the *Journal of Experimental Education.*

If the student wishes to become acquainted with original investigations in the field of educational psychology, for example, he may examine articles published in the *Journal of Educational Psychology;* if he is interested in elementary education, he may consult the *Elementary School Journal.* Such journals report not only the results of original investigations, but suggest opportunities for further research. One of the best ways for the student to locate problems is to read critically the methods and results of investigations reported in periodical literature. The *Psychological Abstracts* and the *Review of Educational Research* will be helpful in ascertaining what has been done in the various fields of study.

Drawing upon one's own experiences. Many students select problems from personal educational experiences or from those in which they become professionally interested. Some as teachers or as students may have experienced difficulties which make them critical of certain educational practices. For example, they may have become aware of need for improvement in teaching procedures. They may wish to make changes in practice in order to obtain improvement in pupil achievement. As superintendents, principals, or supervisors they may have sensed the existence of educational problems concerned with administrative practices or educational policy. Many problems occur to teachers and administrators as a result of general observation of professional activities. In writing a thesis, both teachers and administrators frequently wish to use the opportunity provided by such an activity to study educational problems according to the consensus of research findings or through properly directed experimentation.

Course work as an aid to selecting problems. Each course taken to satisfy requirements for an advanced degree introduces subject matter that should suggest problems. Statements made by the textbook may constitute or form a basis for controversial questions. The student, accordingly, should read textbooks critically—not only in evaluating conclusions reached by an author, but in forming conclusions of his own. Does the author have a sound basis for his conclusions? Are the studies cited in confirmation of his point of view in accord with criteria for sound research? Are his conclusions based upon a limited few observations within a particular school system, or do they represent practices or points of view of many systems? The student may also find many opportunities to question the validity of viewpoints expressed in class discussion. When classes are conducted on a seminar basis, problem-solving attitudes are cultivated; and fellow students frequently present viewpoints that deserve further investigation.

After one tentatively selects a problem, thorough investigation should be made of previous research. One should not make the mistake of surveying only current literature, but, if possible, should explore the literature developed during a period of years. If the student wishes, for example, to investigate effectiveness of classroom incentives, he should search the literature for the purpose of determining the changes that

have occurred during a period of years both in methods of studying the problem as well as in results obtained. There are available in psychology and education a number of reviews which facilitate search for original investigations. Exploration of the literature relating to the tentative problem is one of the most important steps in determining which of its aspects might profitably be explored.

The tentative nature of one's definition of problems. Even after an area or a particular problem has been selected, it must be considered tentative. Seldom are problems as first conceived adequately comprehended or defined with sufficient preciseness to guide one in worthwhile research. Reading, discussion, and reflective thinking all help to define the problem. Prior to the final acceptance of the problem, several questions should be answerea: Has the problem been previously investigated and if so with what results? Are the data needed for study available or readily accessible? If the problem requires new information, are acceptable resources at hand? If the co-operation of others is necessary where may the investigation be conducted? How many subjects are needed, and where may they be obtained? How much will the investigation cost, and how long a time will be needed for its completion?

Selecting a method of research. When one surveys a large number of studies in education one observes certain recurring types of research. Although few studies employ one method exclusively, there may be some justification for thinking of methods under categories such as those described in this volume.

The normative-survey method which has been extensively used in the field of education, is concerned with description of facts and conditions as they exist, without imposition of control upon factors influencing the materials under investigation. The method is essentially one of determining the present status of some educational problems by means of appropriate techniques. Ordinarily one is ultimately interested in ascertaining the adequacy of the status found.

The experimental method of research is used primarily to test and evaluate hypotheses. An experiment is an observation which may be made of controlled, repeated, and varied phenomena. The observer is not only able to control the phenomena under observation, but can produce them at will in order to observe their effects. The influence of experimental factors may be determined by measuring the status of individuals or groups before and after experimental factors have been applied.

Developmental studies and case studies may be applied to individuals, communities, or institutions. Developmental studies follow and record the happenings in some areas of research as they occur; case histories aim at much of the same data from a backward look. Both are concerned with how things come to be and the complex interplay of forces which underlie almost any phenomenon.

The historical method may be defined as that method which de-

scribes a sequence of events during definite chronological periods and how they happen. The historical method includes not only the collection and organization of documentary materials in chronological sequence, but the analysis of causes, conditions, and effects, together with interpretations of their significance for the future.

Foundational research may be an end in itself or applied to bring together the original findings of several, preferably all, of the writers in a given field or problem, to analyze and evaluate them, and to synthesize their conclusions prior to further quantitative research. The aim is to co-ordinate and evaluate the best knowledge in a field relative to some problem and to test hypotheses in the light of already available data. The purpose is primarily to review and synthesize the literature bearing upon a problem.

Regression and correlational research aim to study the "going-togetherness" of different phenomena for purposes of description and prediction. Simple correlation, multiple correlation, and factor analysis are among the worker's more important instruments of research in this area.

Usually one method is more appropriate for attacking a particular problem than another. For example, in justifying increased costs of education to a community or state, it may be desirable to use the historical approach, wherein the investigator might seek to show how additional services have increased the efficiency of education, or the problem may be attacked by the normative-survey method wherein the investigator uses a typical city, county, or state to establish a quantitative relationship between amounts and types of education and efficiency as measured by accepted standards, or the experimental method may be employed to determine whether certain types of educational experience in a particular school may have a measurable effect upon the community in which it is located.

The existence of categories of research methods might lead the student to believe that a given method implies inflexible procedures. This is not the case, for many investigators use several methods in combination. A single method may be chosen with the understanding that other methods may be introduced to supplement and reinforce findings of the major method employed.

Usually one method is used in Masters' theses and in a majority of cases for the doctorate. In the introduction of a thesis, however, there is likely to be some foundational and historical research in order to define the setting of the problem. Throughout all methods there is also a certain amount of reflective thinking or deliberation. Deliberative thinking is particularly important at the beginning of a study where one formulates hypotheses, after data have been presented in the body of the study, and at the end when the investigator reviews the findings and makes observations concerning their applicability.

Tools of research. In the case of a thesis that requires use of a

quantitative method, the tool of research needed for collecting data may be a questionnaire, a rating scale, a test, an inventory, or some other measuring device. All data-gathering devices must be valid, reliable, and discriminating. If a questionnaire is to be used, it will generally need not only to be specially constructed to conform with the investigator's purposes, but particular attention will need to be given to the vocabulary to eliminate ambiguity. A rating scale often must be designed for the particular information to be obtained, but the instrument must be constructed with care if it is to possess reliability and validity. If a test is needed, the investigator may consider numerous tests already available. If a suitable test is not available, he will be obliged to construct one appropriate to his purposes. Prior to using it in an investigation, a preliminary try-out on a representative group of subjects should be made. Tests specifically constructed for an investigation should be subjected to criteria equally rigid as those applied to standardized tests. Chapters III and IV of this volume should be particularly helpful in the choice and construction of data-gathering devices.

Statistical devices. Statistical devices which are sometimes simple and sometimes complex are usually necessary in research. Fully as important as the selection of proper statistical devices is the correct interpretation of results obtained through their use. For the more complex statistical calculations the student should find Chapters VIII, IX, and X helpful. Note carefully the assumptions made in all statistical methods and procedures. Only by giving careful consideration to these assumptions and statistical requirements in the planning of an investigation can the student hope to provide an adequate statistical analysis of the data.

The person who uses qualitative methods must examine information from many sources including quantitative investigations. The student who uses the historical method, for example, frequently must examine masses of factual information presented in statistical form, as well as nonmathematical documents. Although one who uses nonmathematical methods may make no calculations himself, he may find it necessary to interpret the findings of objective studies. As a result of knowledge of statistics one may evaluate a body of quantitative material on which his conclusions must rest. One who conducts a study by means of qualitative methods needs training in quantitative methods; one who develops a study by one of the quantitative methods needs training in methods of qualitative investigation. These two broad categories of research methods are mutually supporting; each makes its own contribution in a comprehensive research program.

Organizing research materials. After the student has collected his materials he is ready to plan the order of presenting his results. Logical organization requires understanding of the various parts of the material collected and their many interrelationships. At this stage the stu-

dent should consider the various parts of a thesis; the statement of the problem, review of previous investigations, the procedure, analysis of data, summary of findings, and recommendations.

The introduction. The introduction describes what the author has tried to do and the procedure used. There is usually a statement of the problem, a review of previous investigations, and a description of the procedure. One usually "warms up" to the subject by indicating the need and value of the study as evidenced by the findings of other investigators in the field. If the student is unable to discover previous related investigations, he should show the relationship between his study and the philosophy underlying his problem or at least indicate its significance. Previous investigations should be reviewed for *method* as well as *content,* and the review should be from the point of view of the problem being studied and not in the nature of a general unoriented summary. The new study may provide opportunity to improve the method or techniques of previous research, to increase the number of subjects, or to refine the tools of research used. Generally, the problem investigated has been explored by someone; only after exploration of previous investigations can the investigator claim intellectual honesty.

The problem of the thesis may be stated and discussed before previous investigations are presented or afterwards. The important consideration is that previous investigations should somewhere be taken into account in outlining the purpose and method of the study. Statement of the problem should include one or more broad objectives followed by two or more specific objectives. Specific objectives serve to clarify the broad objectives and make the purpose of the study more meaningful.

After the problem has been stated, the materials, methods, and scope of the study may be outlined. In the case of quantitative studies statements should be made of the number and kind of subjects used; if a sampling study, how the sample was drawn, instruments of measurement employed, and any other relevant information. The procedure should be so explicit as to enable anyone reading the study to repeat it for purposes of corroboration or refutation of the findings. Description and explanation of procedure in the case of historical and social foundations studies may be even more detailed than in the case of quantitative investigations. A student who uses the historical method should outline his plans for developing the study, indicate sources of data, both original and secondary, and make additional statements that will enable the reader to obtain a clear picture of the procedure followed. In library studies, the number of years covered by the study and the sources from which findings are organized and evaluated should be indicated. If the experimental method is used, it is desirable to describe accurately any apparatus used, as well as the number and kind of subjects, the experimental factors, and methods of experimental control.

Difference of opinion exists with respect to the labeling of the intro-

ductory chapters. Some authorities believe that the introduction should constitute a single chapter of a thesis. Many persons prefer separate chapters for the statement of the problem, review of previous investigations, and outline of investigational procedure. In any case, the introductory orientation is important.

Presentation and analysis of the data. The findings or results of the study may be presented in one or several chapters according to their length and complexity. The nature of the report varies with the type of thesis. Historical and foundational theses frequently employ a number of chapters divided according to the subject discussed. Because of the restricted character of most quantitative studies, a relatively short statement of results may be adequate. In theses which use only one chapter for presenting results, a summary statement at the end of the chapter is usually unnecessary. When the body of the thesis contains several chapters, a short summary at the end of each chapter is appropriate.

In the body of the report the investigator organizes his materials systematically and presents closely related data grouped in sections under appropriate headings. In his arrangement he should be conscious of some systematic plan: (1) presentation of data, (2) interpretation, and (3) indication of meaning and application. In making interpretations and applications, the student may use studies of other investigators as a basis of comparison or corroboration of his results.

There are three modes of presenting results in a thesis, namely, *textual, tabular,* and *graphic.* The investigator who presents his results in the form of explanation, description, or narration uses the textual form. The predominant mode of presentation in qualitative theses is textual. In a thesis employing quantitative data an important part of the presentation of results will be the construction and arrangement of tables and graphs. Tables must be complete and meaningful in and of themselves without further textual explanation. Each table should have a number and title. Graphs clarify and illustrate status facts, trends, and relationships, and are needed in cases in which tables do not reveal important meanings. Tables may be presented in the absence of graphs, but graphs should not be presented without inclusion of data upon which they are based.

In theses employing the historical method, tables may be used to summarize basic facts and to indicate trends.

General summary, conclusions, and recommendations. The third part of a thesis attempts to summarize in a final chapter the major facts obtained together with their interpretations and implications. The usual procedure is to present first a summary of important facts developed, followed by conclusions, implications and recommendations. This part may also include reference to limitations of the study and offer suggestions for further research.

This final chapter provides the busy reader a résumé of the study and affords the investigator an opportunity to take a telescopic view of his findings. He also has opportunity to present implications, interpre-

tations, and applications which relate to some broader problem in education. Here the investigator may proceed beyond his specific findings and consider the broad issues involved.

This chapter may be short or long according to the length and complexity of the body of the report. If several chapters are used in the presentation and analysis of data, the final chapter will probably be longer than in a thesis in which this part contains only one chapter. It may also vary in nature and length with the type of thesis. In an historical study it may, for example, appropriately include a section presenting "future outlook." In a philosophical study one may present important conclusions followed by criticism of method and contribution of previous studies, observations on application of findings, and suggestions for further research.

The abstract. Many institutions require abstracts that may or may not be published. An abstract ordinarily consists of three parts which often are embodied in a few brief paragraphs. The first part describes what the author tried to do or states the problem investigated. In the second part he outlines his procedures. The third part describes the author's findings and states his conclusions and recommendations. Most institutions impose a definite word limitation upon the length of the abstract.

Preparation of thesis for publication. In addition to providing the student valuable training, the thesis should be a contribution to education; and it is often desirable to make it available for publication. The principal criteria for considering whether to submit one's thesis for publication include: (1) Is the study unique in the sense of making a contribution either to method or content? (2) Does it clarify a controversial subject or add knowledge to it? (3) Does it pertain to a current timely topic?

The journals afford the most promising possibilities for publishing theses. The student should study the purposes and nature of various educational and psychological journals in order to determine those that are directly related to his level and field of investigation. The purpose of each journal is ordinarily stated some place in the publication or information can be obtained from the publishers. Some journals sponsor publication of comprehensive studies in the form of monographs.

If the study is to be published in a journal, it must be made as brief as possible. In general, it should contain a statement of the problem and its setting in the literature, the procedure followed, and the results and conclusions. The condensed article usually should not consist of more than twelve or fifteen double-spaced pages. The material of most Masters' theses may be compressed into a single article. Doctors' theses, however, often require several articles if published in periodical form. Relatively comprehensive studies often may be advantageously divided among several important topics.

Editors of journals and other publishing agencies usually have their

own peculiar methods of annotation. It is desirable to study their format prior to submitting an article and to conform as closely as possible to usages illustrated. Note the differences in bibliographical form used by the American Psychological Association publications, the *Review of Educational Research,* and the *Journal of Educational Research.* In all cases, the article should be carefully edited with due regard to clarity, conciseness, rules of grammar, and neatness. The University of Chicago Press *Manual of Style* will be found invaluable in this respect. The author should edit and revise his material several times in order to be sure that it meets publication standards.

APPENDIX B

References

CHAPTER I

Corey, Stephen. "Fundamental research: action research and educational practice," *Amer. Educ. Res. Bull.* (1949).

Fattu, Nicholas A. "Common sense versus experimental inference in educational research," *Amer. Educ. Res. Bull.* (1949).

Hunt, J. McV. "A social agency as a setting for research—the institute of welfare research," *J. of Consult. Psych.* (1949, *13*:69–81).

Johnson, Loaz W. "What administrators want and will use from research workers," *Amer. Educ. Res. Bull.* (1949).

Morrison, J. C. "The accomplishments and promise of research," *Amer. Educ. Res. Bull.* (1949).

Ridenour, Louis N., Jr. "Educational research and technological change," *Amer. Educ. Res. Assoc. Bull.* (1950).

Wrightstone, Wayne. "Evaluation of the experiment with the activity program in the New York City elementary schools," *J. Educ. Res.* (1945, *38*:691–696).

CHAPTER II

Barr, A. S., Burton, W. R., and Brueckner, L. J. *Supervision: Democratic Leadership in the Improvement of Learning* (New York: Appleton-Century-Crofts, 1947, 2nd Ed.).

Cattell, R. B. *Description and Measurement of Personality* (New York: World Book Company, 1946).

Conrad, H. S. "Investigating and appraising intelligence and other aptitudes" (in) *Methods of Psychology.* (Wiley, 1948, pp. 498–538.)

Davis, Frederick B. "Fundamental factors of comprehension in reading," *Psychometrika* (1944, *9*:186).

Edwards, N., and Richey, H. G. *The School in the American Social Order* (Houghton-Mifflin Company, 1947).

Hartman, George W. "How can research help to determine what ought to be," *Amer. Educ. Res. Bull.* (1949).

N. E. A. Educational Policies Commission. *Policies for Education in American Democracy* (Washington, D. C., 1946).

N. S. S. E. 46th Yearbook, *Science Education in American Schools* (Part I, Chicago: University of Chicago Press, 1947).

N. S. S. E. 44th Yearbook, *American Education in the Postwar Period* (Part I, Chicago: University of Chicago Press, 1945).

Peters, C. C. "An experiment with democratized education," *J. Educ. Res.* (1943, *37*:95–99).

Peters, C. C. *Teaching High School History and Social Studies for Citizenship Training* (Coral Gables, Fla.: University of Miami, 1948).

Popham, E., and Place, I. "Aims of college typewriting," *J. Bus. Educ.* (1947, *27*:17–18; *27*:15–16).

Proffitt, M. M., and others. "The measurement of understanding in industrial arts," N.S.S.E. 45th Yearbook, *The Measurement of Understanding* (Part I, Chicago: University of Chicago Press, 1946).

Smith, E. R., Tyler, R. W., and others. *Appraising and Recording Student Progress* (Harper, 1942).

Stiles, L. J., and Dorsey, M. F. *Democratic Teaching in Secondary Schools* (Lippincott, 1950).

CHAPTER III

Baker, G., and Peatman, J. G. "Tests used in Veterans Administration advisement units," *Amer. Psychologist* (1947, *2*:99–102).

Barr, A. S. "The measurement and prediction of teaching efficiency: a summary of investigations," *J. Exp. Educ.* (1948, *16*:1–283).

Buros, Oscar K. (ed). *The Third Mental Measurements Yearbook* (New Brunswick, N. J.: Rutgers University Press, 1949).

Cureton, T. K., and others. "The measurement of understanding in physical education," N.S.S.E. 45th Yearbook, *The Measurement of Understanding* (Part I, Chicago: University of Chicago Press, 1946).

Cronbach, L. J. *Psychological Testing* (Harpers, 1949).

Gessell, A., and Ames, L. B. "The development of handedness," *J. Genet. Psych.* (1947, *7*:155–175).

Guthrie, E. R. "The evaluation of teaching," *Educ. Record* (1949, *30*:109–115).

Lindquist, E. F. (ed). *Educational Measurement* (Washington, D. C.: American Council on Education, 1951).

Lundberg, Donald E. "A simple rating device," *Personnel J.* (1947, *25*:267–270).

McNemar, Q. "Opinion-attitude methodology," *Psych. Bull.* (1946, *44*:289–374).

Oakes, M. E. *Children's Explanations of Natural Phenomena* (Teachers College Contributions to Education, No. 926, 1947).

Thorndike, R. L. *Personnel Selection: Test and Measurement Techniques* (Wiley, 1949).
Thorndike, R. L. (ed.) *Research Problems and Techniques* (Washington, D. C.: U. S. Government Printing Office, Aviation Psychology Research Program, 1947).
"Problems of quantification and objectives in personality areas," Symposium in *Personnel J.* (1948, *17*:141–185).

CHAPTER IV

Adkins, D. C., et al. *Construction and Analysis of Achievement Tests* (Washington, D. C.: U. S. Government Printing Office, 1947).
Adkins, D. C. "Needed research on examining devices," *Amer. Psychologist* (1948, *3*:104–106).
Bechtoldt, H. P., Mauker, J. W., and Stuit, D. B. "The use of order of merit rankings," in *New Methods in Applied Psychology* (College Park: University of Maryland, 1947, pp. 26–33).
Buros, O. K. (ed.). *The Third Mental Measurements Yearbook* (New Brunswick: Rutgers University Press, 1949).
Cronbach, L. J. *Essentials of Psychological Testing* (Harper, 1949).
Davis, F. B. *Item Analysis Data: Their Computation, Interpretation and Use in Test Construction* (Cambridge: Harvard Graduate School of Education, 1946).
Duker, S. "The questionnaire is questionable," *J. Educ. Res.* (1946, *39*:380–383).
Edgerton, Harold A., and others. "Objective differences among various types of respondents to a mailed questionnaire," *Amer. Soc. Review* (1947, *12*:435–444).
Ellis, Albert, and Conrad, Herbert. "The validity of personality inventories in military practice," *Psych. Bull.* (1948, *45*:385–426).
Ellis, Albert. "The validity of personality questionnaires," *Psych. Bull.* (1946, *43*:385–440).
Ellis, Albert. "Personality questionnaires," *Rev. Educ. Res.* (1947, *17*:53–63).
Gulliksen, Harold. *Theory of Mental Tests* (New York: Wiley, 1950).
Lindquist, E. F. *Educational Measurement* (Washington, D. C.: American Council on Education, 1951).
Lundquist, Edward A., and Bittner, R. H. "Using ratings to validate personnel instruments," *Personnel Psych.* (1948, *1*:163–183).
Thorndike, R. L. *Personnel Selection: Test and Measurements Techniques* (New York: Wiley, 1949).

CHAPTER V

STATISTICAL METHODS:

Johnson, Palmer O. *Statistical Methods in Research* (New York: Prentice-Hall, 1949).

Peatman, John G. *Descriptive and Sampling Statistics* (New York: Harper, 1947).

Walker, Helen M. *Elementary Statistical Methods* (New York: Holt, 1943).

MEASUREMENTS:

Buros, O. K. *The Third Mental Measurement Yearbook* (New Brunswick: Rutgers University Press, 1949).

Cronbach, Lee J. *Essentials of Psychological Testing* (New York: Harper, 1949).

Lindquist, E. F., and others. *Educational Measurement* (Washington, D. C.: American Council on Education, 1951).

Thorndike, Robert L. *Personnel Selection: Test and Measurement Techniques* (New York: Wiley, 1949).

NONMATHEMATICAL STATUS STUDIES:

Anderson, Harold H., and Brewer, Joseph E. *Studies of Teachers' Classroom Personalities, II. Effects of Teachers' Dominative and Integrative Contacts on Children's Classroom Behavior. Applied Psychology Monographs,* No. 8 (Stanford University Press, 1946).

Anderson, Harold H., Brewer, Joseph E., and Reed, Mary F. *Studies of Teachers' Personalities, III. Follow-up Studies of the Effects of Dominative and Integrative Contacts on Children's Behavior. Applied Psychology Monographs,* No. 11, (Stanford University Press, 1946).

Brueckner, Leo J. "Learning and meaning of democracy through participation, observation, and study," *Nat'l Elem. Principal* (1948, *27*:39–43).

Lingren, Vernon C. "Criteria for the evaluation of in-service activities in teacher education," *J. Educ. Res.* (1948, *42*:62–68).

Snyder, William U. "The present status of psychotherapeutic counseling," *Psych. Bull.* (1947, *49*:297–386).

MATHEMATICAL STATUS STUDIES: NONVARIATE:

Blanchard, B. Everard, "A social acceptance study of transported and non-transported pupils in a rural secondary school," *J. Exper. Educ.* (1947, *15*:291–303).

Bonney, Merl E. "A study of the sociometric process among sixth-grade children," *J. Educ. Psych.* (1946, *37*:359–72).

Morton, John A. "A study of children's mathematical questions as a clue to grade placement of arithmetic topics," *J. Educ. Psych.* (1946, *37*:293–315).

Otto, Henry J. *Organizational and Administrative Practices in Elementary Schools in the United States* (No. 4544. Austin, Texas: The University of Texas Press, 1945).

MATHEMATICAL STATUS STUDIES: VARIATE:

Hopkins, J. W. "Height and weight of Ottawa elementary school children in two socio-economic strata," *Human Biology* (1947, *19*:68–82).

Jones, Harold E. "The sexual maturing of girls as related to growth in strength," *Res. Quart. Amer. Assoc. Health, Phys. Educ., and Recreat.* (1947, *18*:135–143).

Pierce, Truman M. *Controllable Community Characteristics Related to the Quality of Education* (New York: Bureau of Publications, Teachers College, Columbia University, 1947).

Wood, Ben D., and staff. *1949 Fall Testing Program in Independent Schools and Supplementary Studies* (Educational Records Bulletin, No. 53. New York: Educational Records Bureau, 1950).

SOME RECENT SCHOOL SURVEYS:

Brewton, John E. (director). *Public Education in New Mexico* (A report of the New Mexico Educational Survey Board. Nashville, Tennessee: George Peabody College for Teachers, 1948).

Chase, Francis S., and Morphet, Edgar L. (directors). *The Forty-Eight State School Systems* (Chicago: Council of State Governments, 1949).

Cherry, Ralph W. (director). *Public Education in Harlan County, Kentucky* (Bulletin of the Bureau of School Service, Vol. 19, No. 2. Lexington: University of Kentucky, 1946).

North Carolina State Education Commission. *Education in North Carolina Today and Tomorrow* (Raleigh: The Commission, 1948).

Seagers, Paul W., and Holmstedt, Raleigh W. (directors). *A Comprehensive Co-operative Study of the Schools of Perry Township, Marion County, Indiana* (School Survey Series, No. 4. Bloomington: Indiana University, 1949).

NOT READILY ASSIGNED TO ANY ONE TYPE:

Emans, Lester M. "In-service education through co-operative curriculum study," *J. Educ. Res.* (1948, *41*:695–702).

Gesell, Arnold; Illg, Frances L., Ames, Louise B., and Bullis, Grace E. *The Child from Five to Ten* (New York: Harper, 1946).

Pierce, Truman Mitchell. *Controllable Community Characteristics Related to the Quality of Education* (New York: Teachers College, Columbia University, 1947).

CHAPTER VI

Anderson, Kenneth E. "A frontal attack on the basic problem in evaluation: the achievement of instruction in specific areas," *J. Experimental Educ.* (1950, *18*:163–174).

Cornell, Francis G. "Sample plan for a survey of higher education enrollment," *J. Experimental Educ.* (1947, *15*:213–218).

Deming, William E. *Some Theory of Sampling* (New York: John Wiley and Sons, 1950).

Hansen, Morris H., and Hurwitz, William N. "The problem of non-response in sample surveys," *J. Amer. Statistical Assoc.* (1946, *41*:517–529).

Harris, Marilyn, Howitz, D. G., and Mood, A. M. "On the determination of sample sizes in designing experiments," *J. Amer. Statistical Assoc.* (1948, *43*:391–402).

Hartkemeier, H. P. *Principles of Punch-Card Machine Operation* (New York: Thomas Y. Crowell, 1942).

Johnson, Palmer O. *Statistical Methods in Research* (Chapter 9. New York: Prentice-Hall, 1949).

Marks, Eli S. "Sampling in the revision of the Stanford-Binet Scale," *Psych. Bull.* (1947, *44*:413–434).

Marks, Eli S. "Some sampling problems in educational research," *J. Educ. Psych.* (1951, *42*:85–95).

Reid, Seerly, "Respondents and nonrespondents to mail questionnaires," *Educ. Res. Bull.* (1942, *21*:87–96).

Yates, Frank, *Sampling Methods for Censuses and Surveys* (London: Charles Griffin and Co., 1949).

CHAPTER VII

CASE STUDIES AND CLINICAL DIAGNOSIS:

Barr, A. S., Burton, W. H., Brueckner, Leo J. *Supervision: Democratic Leadership in the Improvement of Learning* (New York: Appleton-Century-Crofts, Inc., 1947).

Bell, John E. *Projective Techniques* (New York: Longmans, 1948).

Burton, Arthur, and Harris, Robert E. (editors). *Case Histories in Clinical and Abnormal Psychology* (New York: Harper, 1947).

Cronbach, Lee J. *Essentials of Psychological Testing* (New York: Harper, 1949).

McKinney, Fred, "Case history norms for unselected students and students with emotional problems," *J. Consulting Psych.* (1947, *11*:258–69).

Muench, George A. "An evaluation of nondirective psychotherapy by means of the Rorschach and other indices." *Applied Psych. Monographs* (No. 13., 1947).

Strang, Ruth, *Counseling Technics in College and Secondary School* (Revised and enlarged edition. New York: Harper, 1949).

Thurstone, L. L. *The Dimensions of Temperament* (Report No. 42, Psychometric Laboratory. Chicago: The University of Chicago Press, 1947).

Zubin, Joseph, "Recent advances in screening the emotionally maladjusted," *J. Clinical Psych.* (1948, *4*:56–63).

COMPARATIVE STUDIES:

A. Those growing out of the application of the logical principles of agreement and double agreement:

Ankerman, Robert C. "Differences in the reading status of good and poor eleventh-grade students," *J. Educ. Res.* (1948, *41*:498–515).

Barr, A. S. *Characteristic Differences in the Teaching Performance of Good and Poor Teachers of the Social Studies* (Bloomington, Illinois: Public School Publishing Co., 1929).

Birkeness, Valborg, and Johnson, Harry C. "A comparative study of delinquent and non-delinquent adolescents," *J. Educ. Res.* (1949, *42*:561–572).

Heisler, Florence, "A comparison of comic book and non-comic book readers of the elementary school," *J. Educ. Res.* (1947, *40*:458–464).

Orr, Harriet Knight, "A comparison of the records made in college by students from full accredited high schools with those of students having equivalent ability, from second- and third-class high schools," *J. Educ. Res.* (1949, *42*:353–364).

Ramharter, Hazel K., and Johnson, Harry C. "Methods of attack used by good and poor achievers in attempting to correct errors in six types of subtraction involving fractions," *J. Educ. Res.* (1949, *42*:586–597).

Van Dalen, D. B. "A differential analysis of the play of adolescent boys," *J. Educ. Res.* (1947, *41*:204–213).

COMPARATIVE STUDIES:

Ebaugh, C. D. *Education in Peru* (U. S. Office of Education, Bulletin No. 3, 1946).

Kandel, I. L. "National backgrounds of education," *Twenty-fifth Yearbook, National Society of College Teachers of Education* (Chicago: University of Chicago Press, 1937).

Lewis, Gertrude Weiss (editor) *Resolving Social Conflicts: Selected Papers on Group Dynamics* (New York: Harper, 1948).

Northrop, F. S. C. *The Meeting of East and West: An Inquiry Concerning World Understanding* (New York: Macmillan, 1946).

HISTORICAL STUDIES:

Brubacher, J. S. *A History of the Problems of Education* (New York: McGraw-Hill, 1947).

Butts, R. F. *A Cultural History of Education* (New York: McGraw-Hill, 1947).

Collingwood, R. G. *The Idea of History* (New York: Oxford University Press, 1946).

Curti, Merle (chairman). *Theory and Practice in Historical Study* (New York: Social Science Research Council, 1946).

Edwards, Newton, and Richey, H. G. *The School in the American Social Order* (Boston: Houghton Mifflin, 1947).

Good, H. G. "Current Historiography in education," *Review of Educ. Res.* (1949, *9*:456–59).
Hockett, H. C. *Introduction to Research in American History* (New York: Macmillan, 1948).
Moehlman, Arthur H. "Toward a new history of education," *School and Society* (1946, *43*:57–60).
Reisner, Edward H. "The more effective use of historical background in the study of education," *The Uses of Background in the Interpretation of Educational Issues* (Yearbook 25. Ann Arbor, Michigan: Ann Arbor Press, 1944).
Reisner, Edward, Kandel, I. L., and Knight, Edgar W. "New emphases in history of education in response to war and post-war demands," *National Society of College Teachers of Education* (Yearbook 29. Ann Arbor, Michigan: Ann Arbor Press, 1944).
Theory and Practice in Historical Study: A Report of the Committee on Historiography (Bulletin 54. New York: Social Science Research Council, 1946).
Ulich, Robert, *Three Thousand Years of Educational Wisdom* (Cambridge, Massachusetts: Harvard University Press, 1947).

CHAPTER VIII

Brunswik, Egon, "Systematic and representative design of psychological experiments with results in physical and social perception," *Proceedings of the Berkeley Symposium* (Univ. Calif. Press, 1949, 143–207).
Churchman, C. West, *Theory of Experimental Inference* (New York: Macmillan Co., 1948).
Cochran, William C., and Cox, Gertrude M. *Experimental Designs* (New York: John Wiley and Sons, 1950).
Engelhart, Max D. "Suggestions with respect to experimentation under school conditions," *J. Experimental Educ.* (1946, *14*:225–44).
Fisher, Ronald A. *The Design of Experiments* (4th edition. Edinburgh: Oliver and Boyd, Ltd., 1947).
Freeburne, Cecil M. "The influence of training in perceptual span and perceptual speed upon reading ability," *J. Educ. Psych.* (1949, *40*: 321–352).
Heidgerken, Loretta E. "An experimental study to measure the contribution of motion pictures and slidefilms to learning certain units in the course introduction to nursing arts," *J. Experimental Educ.* (1948, *17*:261–81).
Johnson, Donovan A. "An experimental study of the effectiveness of film strips in teaching geometry," *J. Experimental Educ.* (1949, *17*: 363–72).
Johnson, Palmer O. "Modern statistical science and its function in educational and psychological research," *Scientific Monthly* (1951, *62*:385–396).

Lindquist, E. F. *Statistical Analysis in Educational Research* (Boston: Houghton Mifflin Co., 1940).

Von Eschen, Clarence R. "The improvability of teachers in service," *J. Experimental Educ.* (1945, *14*:135–56).

CHAPTER IX

American Council on Education, *Exploring Individual Differences.* A report of the 1947 invitational conference on testing problems. (Washington, D. C., 1948).

American Council on Education (E. F. Lindquist editor). *Educational Measurement* (Washington, D. C.: 1951, Chapters 6 and 7).

Barr, A. S., et al. "The prediction of teaching efficiency," *J. Experimental Educ.* (1948, *15:*).

Donahue, Wilma T., Coombs H., and Travers, R.M.W. (editors). *Measurement of Student Adjustment and Advancement* (Ann Arbor: University of Michigan Press, 1949).

Detchen, Lily, "Effect of a measure of interest factors on the prediction of performance in a college social science comprehensive examination," *J. Educ. Psych.* (1946, *37*:45–52).

Educational Testing Service, "Validity, norms, and the verbal factor," *Proceedings of the 1948 Invitational Conference on Testing Problems* (New York: 1948).

Gulliksen, Harold, *Theory of Mental Tests* (New York: John Wiley and Sons, 1950).

Jackson, Robert, "The selection of students for freshman chemistry by means of discriminant functions," *J. Experimental Educ.* (1950, *18:* 209–214).

Johnson, Palmer O. *Statistical Methods in Research* (New York: Prentice-Hall, 1949).

Keys, Noel, "The value of group test I.Q.'s for prediction of progress beyond high schools," *J. Educ. Psych.* (1940, *31*:81–93).

Patterson, C. H. "On the problem of the criterion in prediction studies," *J. Consulting Psych.* (1946, *10*:277–80).

CHAPTER X

American Educational Research Association, *Psychological Tests and Their Uses Review Educ. Res.* (1947, *17:* Chapters 2 and 7).

Bittner, Reigh H., and Wilder, Carlton E. "Expectancy tables: a method of interpreting correlation coefficients," *J. Experimental Educ.* (1949, *14*:245–52).

Fisher, R. A. "The analysis of covariance method for the relation between a part and the whole," *Biometrics* (1947, *3*:65–68).

Gulliksen, Harold, *Theory of Mental Tests* (New York: John Wiley and Sons, 1950).

Honzik, M. P., Macfarlane, J. W., and Allen I. "The stability of

mental test performance between two and eighteen years," *J. Experimental Educ.* (1948, *17*:309–15).
Jackson, Robert W. B. "Some pitfalls in statistical analysis of data expressed in form of I. Q. scores," *J. Educ. Psych.* (1940, *31*:677–85).
Johnson, Palmer O. *Statistical Methods in Research* (New York: Prentice-Hall, 1949).
Stroud, J. B. "Rate of visual perception as a factor in rate of reading," *J. Educ. Psych.* (1945, *36*:487–98).
Thurstone, L. L. *Multiple-Factor Analysis* (Chicago: The University of Chicago Press, 1947).
Turnbull, William W. "The relationship between verbal factor scores and other variables," *Proceedings of the 1948 Invitational Conference on Testing Problems* (Princeton, N. J.: Educational Testing Service, 1948).

CHAPTER XI

Dewey, John. *Sources of a Science of Education* (New York: Horace Liveright, 1929).
"Educational implications of population changes," *Review Educ. Res.* (1946, *16*:50–55).
Edwards, Newton. "Educational implications of population change," *Educ. Forum* (1946, *10*:281–288).
Good, Carter V., Barr, A. S., and Scates, Douglas. *Methodology of Educational Research* (New York: Appleton-Century-Crofts, 1936).
Jones, Harold E. *Development in Adolescence* (New York: Appleton-Century-Crofts, 1943).
La Duke, Charles V. "The measurement of teaching ability," *J. Experimental Educ.* (1945, *14*:75–100).
Lins, Leo Joseph, "The prediction of teaching efficiency," *J. Experimental Educ.* (1946, *15*:2–60).
Ragsdale, C. E., and others, "Adventures in rural education; a three year report," *J. Experimental Educ.* (1944, *12*:245–348).
Rostker, L. E. "The measurement of teaching ability," *J. Experimental Educ.* (1945, *14*:6–51).
Westaway, F. W. *Scientific Method: Its Philosophical Basis and Modes of Application* (London: Blackie and Son, 1937).

INDEX OF AUTHORS

INDEX OF SUBJECTS